Newcastle University

ePrints

Cumpson PJ, Sano N, Fletcher IW, Portoles JF, Bravo-Sanchez M, Barlow AJ. Multivariate analysis of extremely large ToFSIMS imaging datasets by a rapid PCA method. *Surface and Interface Analysis* 2015, 47(10), 986-993.

Copyright:

© 2015 The Authors. Surface and Interface Analysis published by John Wiley & Sons Ltd.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI link to article:

http://dx.doi.org/10.1002/sia.5800

Date deposited:

17/09/2015

This work is licensed under a Creative Commons Attribution 4.0 International License

Newcastle University ePrints - eprint.ncl.ac.uk

Rapid communication

SURFACE and INTERFACE ANALYSIS

Received: 11 May 2015 Revised: 22 June 2015 Accepted: 23 June 2015 Published online in Wiley Online Library: 27 August 2015

(wileyonlinelibrary.com) DOI 10.1002/sia.5800

Multivariate analysis of extremely large ToFSIMS imaging datasets by a rapid PCA method

Peter J. Cumpson,* Naoko Sano, Ian W. Fletcher, Jose F. Portoles, Mariela Bravo-Sanchez and Anders J. Barlow

Principal component analysis (PCA) and other multivariate analysis methods have been used increasingly to analyse and understand depth profiles in X-ray photoelectron spectroscopy (XPS), Auger electron spectroscopy (AES) and secondary ion mass spectrometry (SIMS). These methods have proved equally useful in fundamental studies as in applied work where speed of interpretation is very valuable. Until now these methods have been difficult to apply to very large datasets such as spectra associated with 2D images or 3D depth-profiles. Existing algorithms for computing PCA matrices have been either too slow or demanded more memory than is available on desktop PCs. This often forces analysts to 'bin' spectra on much more coarse a grid than they would like, perhaps even to unity mass bins even though much higher resolution is available, or select only part of an image for PCA analysis, even though PCA of the full data would be preferred.

We apply the new 'random vectors' method of singular value decomposition proposed by Halko and co-authors to time-of-flight (ToF)SIMS data for the first time. This increases the speed of calculation by a factor of several hundred, making PCA of these datasets practical on desktop PCs for the first time. For large images or 3D depth profiles we have implemented a version of this algorithm which minimises memory needs, so that even datasets too large to store in memory can be processed into PCA results on an ordinary PC with a few gigabytes of memory in a few hours. We present results from ToFSIMS imaging of a citrate crystal and a basalt rock sample, the largest of which is 134GB in file size corresponding to 67 111 mass values at each of 512 × 512 pixels. This was processed into 100 PCA components in six hours on a conventional Windows desktop PC. © 2015 The Authors. *Surface and Interface Analysis* published by John Wiley & Sons Ltd.

Keywords: SIMS; principal component analysis; PCA; multivariate analysis; time-of-flight secondary ion mass spectrometry; data analysis

Introduction

Principal component analysis (PCA)[1] is a powerful tool for surface analysis data and has many applications. It can provide an overview of exactly the type of complex data that modern surface analysis instruments produce. PCA can be used for revealing relations between spectra and spatial position, detecting outliers and finding patterns in massive datasets that are otherwise impossible to study by simply plotting one variable against another, amongst other applications. PCA has therefore been an important method of analysing spectra and images in surface analysis for at least the last 25 years. There exist excellent examples[2–8] and analytical reviews[9–13] of its use in the literature, applied to a range of problems. At the core of PCA software is Singular Value Decomposition (SVD), a matrix algebra method for decomposing spectra into orthogonal (i.e. independent) components.[14,15] Consider the simple case of an image containing just four pixels, as shown in Fig. 1. We have labelled these pixels a, b, c and d. We think of pixels (such as those taken by a digital camera for example) as containing intensity information, or perhaps three numbers representing the intensity of red, green and blue (RGB) light in a digital image. In Auger electron spectroscopy (AES), X-ray photoelectron spectroscopy (XPS) or ToFSIMS we may have a complete spectrum at each pixel, with perhaps somewhere between $n = 100$ and $n = 100\,000$ numbers rather than just three RGB values. Each pixel therefore contains a complete spectrum for a range of electron energy (AES and XPS) or mass-to-charge ratio (ToFSIMS).

Singular Value Decomposition (SVD) is often described in linear algebra or numerical analysis texts whereby we can express a matrix as

$$A = U\,S\,V^T \tag{1}$$

Or, equivalently, in tableau form showing the elements of these matrices;

* *Correspondence to: P. J. Cumpson, National EPSRC XPS User's Service (NEXUS) Laboratory, School of Mechanical and Systems Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom.*
E-mail: peter.cumpson@ncl.ac.uk

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

National EPSRC XPS User's Service (NEXUS) Laboratory, School of Mechanical and Systems Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom

© 2015 The Authors. *Surface and Interface Analysis* published by John Wiley & Sons Ltd.

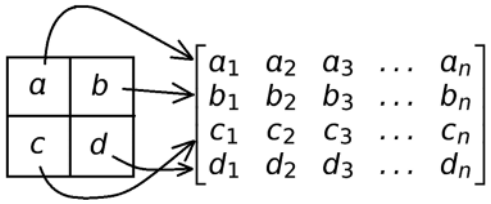


Figure 1. The mass spectrum acquired at each pixel becomes a row of the 'data matrix' for multivariate analysis. The data matrix therefore has the same number of rows as pixels in the image, and the same number of columns as mass values in the spectra.

$$\begin{bmatrix} a_1 & a_2 & a_3 & \dots & a_n \\ b_1 & b_2 & b_3 & \dots & b_n \\ c_1 & c_2 & c_3 & \dots & c_n \\ d_1 & d_2 & d_3 & \dots & d_n \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ u_{21} & u_{22} & u_{23} & u_{24} \\ u_{31} & u_{32} & u_{33} & u_{34} \\ u_{41} & u_{42} & u_{43} & u_{44} \end{bmatrix} \quad (2)$$
$$\times \begin{bmatrix} s_1 & 0 & 0 & \dots & 0 \\ 0 & s_2 & 0 & \dots & 0 \\ 0 & 0 & s_3 & \dots & 0 \\ 0 & 0 & 0 & \dots & s_n \end{bmatrix} \times \begin{bmatrix} v_{11} & v_{21} & v_{31} & \dots & v_{1n} \\ v_{12} & v_{22} & v_{32} & \dots & v_{2n} \\ v_{13} & v_{23} & v_{33} & \dots & v_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ v_{1n} & v_{2n} & v_{3n} & \dots & v_{nn} \end{bmatrix}$$

where **U** and **V** are unitary matrices (rows and columns orthonormal) and **S** is zero everywhere except along its leading diagonal. SVD is therefore a generalisation of matrix eigenvalue decomposition. Many pieces of software are available that, given the data matrix **A**, will return the matrices **U**, **S** and **V**.

The really valuable function of SVD is to break the data into three parts, each of which is much easier to understand;

1. The component parts of the spectra present—these components are separated-out into different columns in the **V** matrix
2. How important each of those components is (organised in decreasing order of importance in the **S** matrix)
3. Where in the image each of those components comes from (separated out into different columns in the **U** matrix).

Therefore while there are plenty of cases in which PCA can be troublesome because of the assumption that the image is made up of a ***linear*** combination of spectra (almost never rigorously the case), nevertheless the value of PCA in elucidating components in a massive amount of data is compelling. Plotting-out the component spectra and the origin of those components within the image can be extremely helpful in understanding surface chemistry.

We should always remember that not all images have an underlying linear relationship suitable for linear multivariate analysis such as this. For example, recently we developed[16] MAFI—Multivariate Auger Feature Imaging—in order to image the graphite-like *versus* diamond-like coordination in novel carbon-based materials, where the carbon Auger feature changes in a nonlinear way as bonding changes from sp^2 to sp^3.

The capability of modern personal computers is extremely good, so that carrying-out PCA calculations is very quick and easy for small-to-moderate data sets. For example we used SVD in the analysis of Angle-Resolved XPS[17] data to overcome ill-conditioning.[18] Because the calculation is so rapid, at NEXUS we perform PCA of sputter depth-profile datasets routinely as one step in understanding the information present, even if the results of this multivariate analysis may not appear in the final publication. However the matrices involved in Angle-Resolved XPS or XPS sputter depth-profiling at a single position are much smaller than those that arise from 2D or 3D imaging. For example, an XPS or AES spectrum containing perhaps 1000 channels and repeated for perhaps 30 levels of a depth-profile leads to a data matrix of 30×1000 elements. SVD is easy and quick for a data matrix of this size. By comparison, modern XPS and SIMS instruments can produce 2D images (or 3D depth-profiles) in which a spectrum is associated with each pixel (or voxel). This leads to enormous datasets, especially in SIMS, where the spectra themselves are larger, and the high lateral resolution of SIMS means the density of pixels (or voxels) may be very high over a similar area of interest. Even in XPS though, the size of 2D and 3D data generated occasionally makes PCA difficult or impractical to apply to the whole dataset by existing methods.

Internally the computer algorithms that are used to solve the SVD problem are complex, and some of them have run-time and memory requirements for processing large imaging spectroscopy datasets that is far beyond what is available even on supercomputers. The computationally intensive step is SVD, and the time required scales much more rapidly than the size of the matrix **A**, so that doubling the number of pixels in an image much more than doubles the calculation time. For imaging datasets in XPS this means calculation times in the range days to weeks, whereas for 3D SIMS data generated by existing state-of-the-art instruments the SVD calculation would be from months to years on a desktop PC using these conventional algorithms, even if enough memory were available. Therefore, while some examples of excellent results from PCA of images exist in the literature (often 'binning' data at low resolution with no other aim than reducing the computational workload) the community is now in a position where PCA applied to depth profiles is routine and valuable, while PCA applied to 2D images (or 3D depth-profiles) is rare but potentially very useful indeed. Some acceleration should be possible by moving to high performance computers or use of Graphical Processor Units (GPUs) to speed-up existing algorithms, but these have not been adopted widely. ToFSIMS datasets are often extremely 'sparse' in the sense of containing many zeros, and some attempts have been made to use this to accelerate the calculation.[19] The purpose of this paper is to investigate how to do SVD (and therefore PCA) in a reasonable time with a typical personal computer for some of the largest data sets being generated by modern surface analysis instruments. In the remainder of this paper we do this by;

(a) Applying a new SVD algorithm (not previously used in surface analysis) to increase the speed of PCA calculation by several orders of magnitude,
(b) Overcoming the memory limit in a typical PC using a variant of the above algorithm that allows most of the data to remain on hard disc during the calculation, so that it is fetched in pieces for processing. This takes longer (sometimes several hours) but allows PCA to be applied to extremely large images on small computers.

Methods (a) and (b) allow us to apply PCA to moderate and large ToFSIMS datasets respectively in a reasonable time (between and few seconds and 12 h or so) on a conventional Windows personal computer.

We do not attempt in this work to take advantage of sparseness, as this adds some complexity to the algorithms involved and can preclude some types of pre-processing of ToFSIMS data, though in the future it may be possible to combine these approaches.

© 2015 The Authors. *Surface and Interface Analysis* published by John Wiley & Sons Ltd.

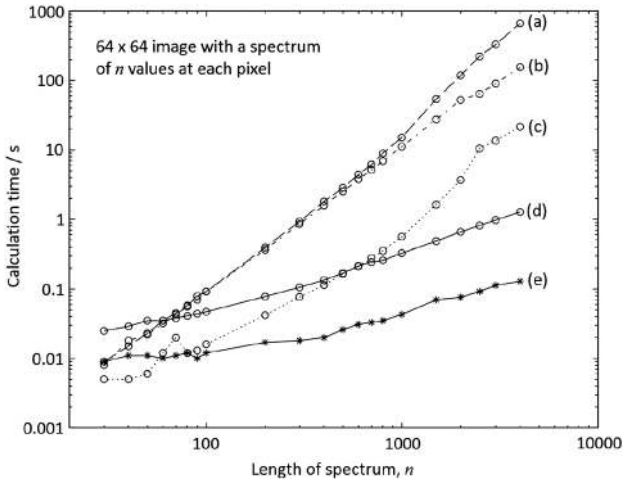


Figure 2. Calculation times recorded for PCA of complete 64 × 64 pixel images having increasing spectra associated with each pixel. Calculations were performed with three different algorithms in two common numerical analysis packages (Octave and Matlab), (a) Octave xGESVD, (b) Octave xGESSD, (c) Matlab xGESVD, (d) Octave, our implementation of Halko *et al.*'s algorithm (RV1), (e) Matlab, our implementation of Halko *et al.*'s algorithm (RV1).

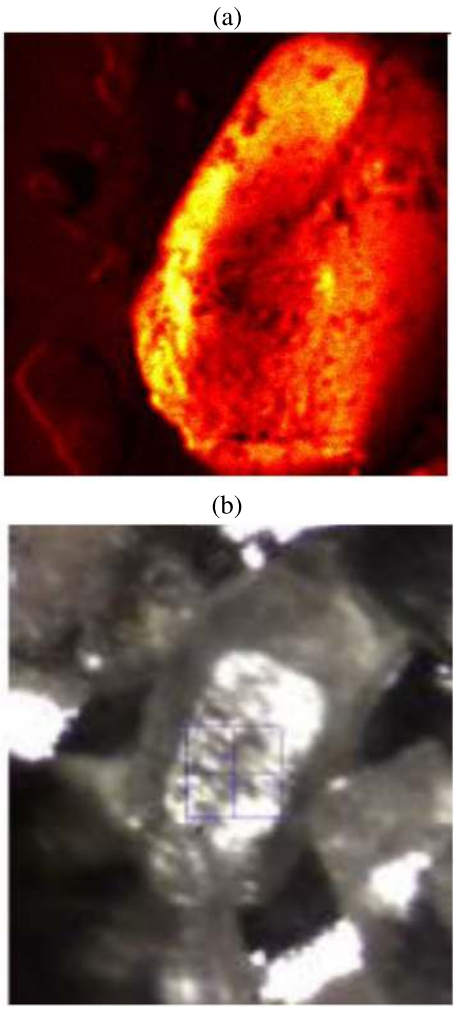


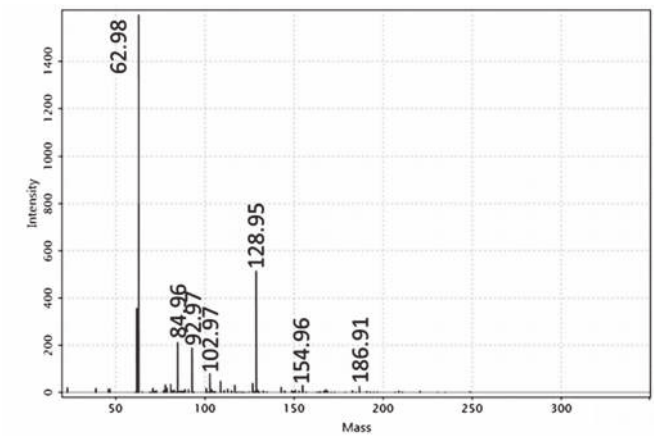
Figure 3. (a) Total ion ToFSIMS image, in positive ion mode, of a sodium citrate crystal. The field of view is 1 mm × 1 mm. Also shown is (b), an optical image of a similar but not the same crystal. These crystals are typically around 0.5 to 1.0 mm in size. This figure is in colour in the online version.

Requirements of surface analysis

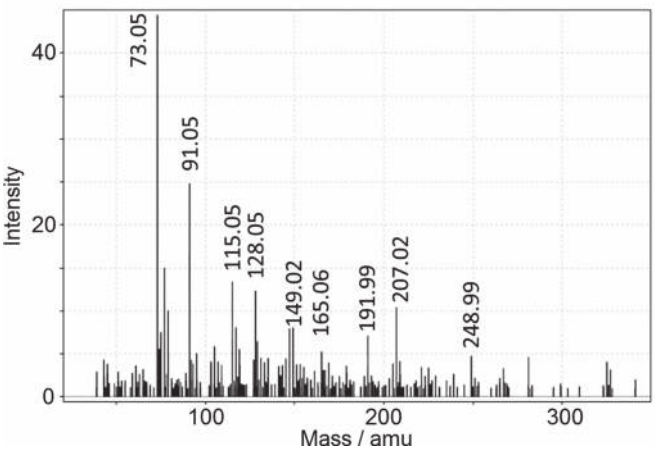
We can identify three key requirements of an algorithm suitable for surface analysis imaging applications;

1. It should be fast for images containing a large number of pixels and having large spectra in those pixels
2. The memory requirements should be within those available on easily-accessible PCs,
3. It should be capable of decomposing low-rank data matrices, i.e. we believe that the spectra in the image are made-up of a small number of factors, almost certainly below 100, and certainly a very small number compared to the total number of pixels.

Note that there is an extensive literature and recommendations on data pre-processing required in XPS and ToFSIMS, which may include normalisation, mean-centering and Poisson distribution variance correction.[20] While numerically quick and easy it is not the purpose of this paper to examine them, as they do not affect the primary issue which is the scale of spectral imaging data. The issues of pre-processing are the same whatever SVD or PCA algorithm is used, and whether the data matrix has 10^2 elements or 10^{10}. However because the choice of preprocessing method depends on the application, and there are publications that discuss



(a) Positive ion spectrum of sodium citrate



(b) Positive ion spectrum of the sodium citrate (double sided tape)

Figure 4. (a) Positive ion spectrum of sodium citrate. (b) Positive ion spectrum of the substrate (double sided tape).

© 2015 The Authors. *Surface and Interface Analysis* Published by John Wiley & Sons Ltd.

SURFACE and INTERFACE ANALYSIS

preprocessing of surface analytical data in depth,[21,22] it will not be considered here.

Four SVD algorithms

We will not describe the details of any SVD algorithms here, but from the point of view of the criteria set out above it is worthwhile highlighting the properties of different classes of algorithm for performing exactly the same SVD function. We list these below in the historical order in which they were developed; we will not explain the terminology and acronyms here, which are included simply to facilitate the reader in looking-up the details of how they work if required. Properly used, each of these algorithms delivers the same results except for small truncation and rounding errors intrinsic to numerical computation, though the time taken and memory required during the calculation can be very different. We will use the terminology of LAPACK[23] in describing them, as this is probably the most widely used numerical subroutine library.

1. Householder reduction to bidiagonal form, followed by QR and shifts[24,25] to eliminate off-diagonal elements. This is implemented in the routine xGESVD in LAPACK. This type of SVD algorithm has been used extensively since the time of its first description in the literature by Golub, Kahan and Reinsch in the period 1965 to 1970.
2. A divide and conquer algorithm[26] (implemented in the routine xGESDD in LAPACK). This is faster for very large matrices than algorithm 1 (so is closer to what we need for surface analytical imaging), but uses more memory. This was described in the literature around 1995 and added to major linear algebra libraries in the late 1990s (e.g. LAPACK V3.0 in 1999).
3. A 'Random vectors' version of the block Lanczos method because of Halko *et al.*,[27] and developed as part of his PhD work.[28] This is a very new method, appearing in the literature around 2010/2011. Variants of the algorithm are still appearing in the primary numerical analysis literature. This method works well for extremely large data matrices, requires that the data matrix be of low rank (an assumption PCA practitioners always must make in any case) and is much faster than algorithm 1 or 2 in this limit. We will call this algorithm 'RV1'

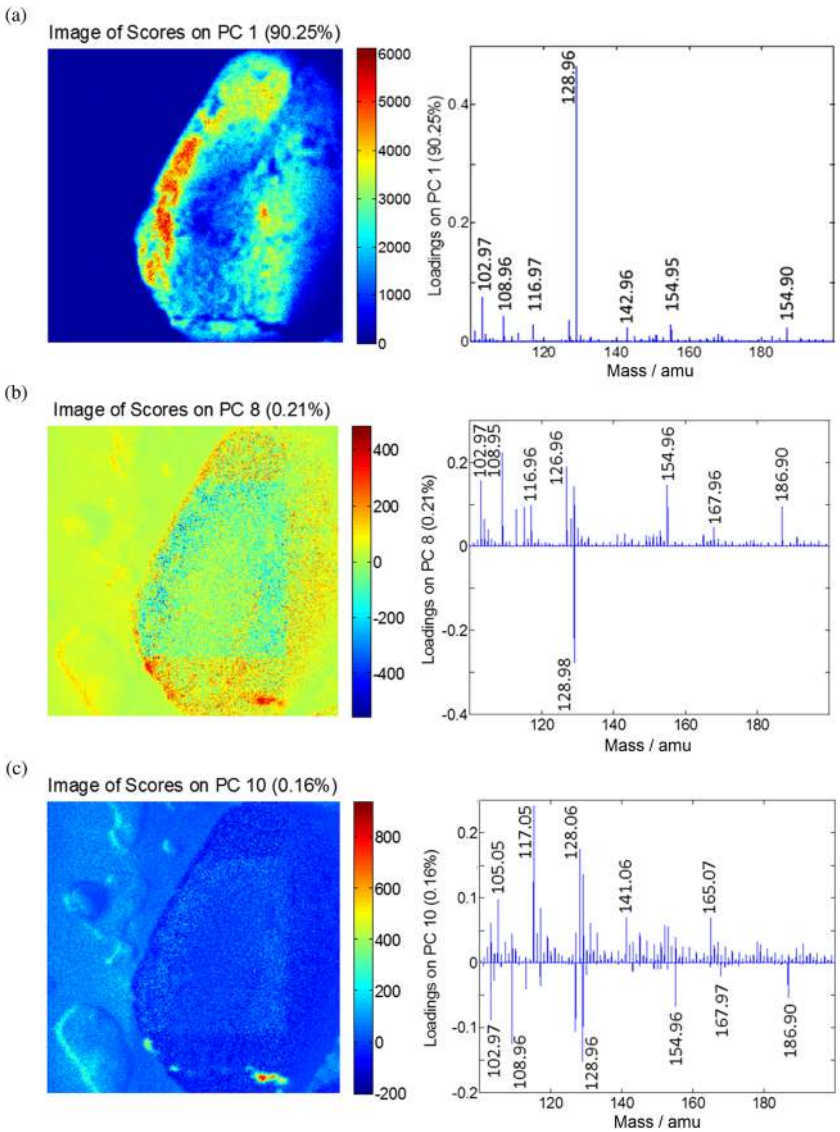


Figure 5. PCA results of sodium citrate. The size of each image component is 1 mm × 1 mm, 256 × 256 pixels.

© 2015 The Authors. *Surface and Interface Analysis* published by John Wiley & Sons Ltd.

4. An 'out-of-core' variant of algorithm 3 has been demonstrated that greatly reduces memory requirements too. This is achieved by organising the numerical steps involved efficiently so that the data matrix is operated-upon in pieces, and therefore the whole data matrix need not be in memory at any one time. We will call this algorithm 'RV2'

RV1 is the most rapid of these algorithms for large image PCA, but relies on enough computer memory being available to hold the PCA data matrix in memory. RV2 is slower because it allows most of the PCA data matrix to reside on disc, loading successive segments of this data matrix into memory for processing. Although slower, RV2 can process very large data matrices, and is quite capable of processing the largest ToFSIMS datasets on an ordinary PC. The largest dataset we have so far processed in this way occupies 134GB when uncompressed, and this processing took only around 6 h.

Methods 1 and 2 (xGESVD and xGESDD) are widely used, and these are the algorithms inside the commonly used software such as Matlab and Octave that perform SVD (and which are therefore inside much PCA software). A good detailed comparison of the first two algorithms can be found in the work of Cline and Dhillon.[29] An excellent, though rather old, animated illustration of how the Golub–Kahan–Reinsch algorithm works is available on Youtube.[30]

No built-in function for RV1 or RV2 exists in Matlab or Octave. Yet algorithm RV1 and RV2 offer the prospect of extending SVD (and hence PCA and other multivariate methods that depend on it) to a much wider range of problems in the very near future, in the same revolutionary way that the Fast Fourier Transform (FFT) extended the use of Fourier methods in digital signal processing in the period 1965 to 1980. XPS and SIMS instruments are now so efficient at providing spectra that one of the clear bottlenecks in delivering results to our collaborators at NEXUS is the analysis of these huge datasets. We have therefore, from scratch, implemented algorithms RV1 and RV2 based on the publications of Halko *et al.* as software within a MATLAB (and Octave) desktop PC environment, and we now have considerable experience of applying these methods to XPS and SIMS data. Similar methods have previously been applied to MALDI and IR

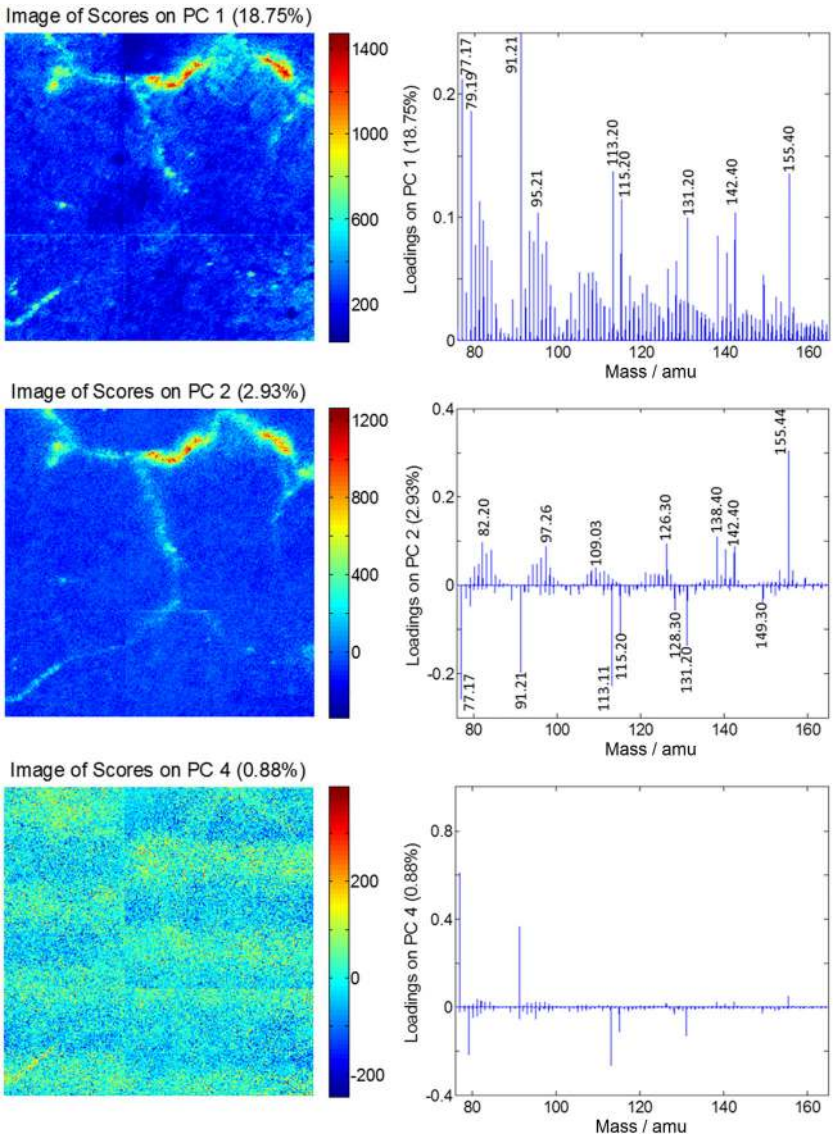


Figure 6. PCA results of the Ontong Java Plateau basalt. The field of view in each component is 1.6 mm × 1.6 mm, 416 × 416 pixels. This region of interest was taken from a larger tiled image, and a horizontal and vertical tile boundary can be seen, particularly in components 1 and 4. Component 4 is an instrumental artefact as described in the text. This figure is in colour in the online version.

© 2015 The Authors. *Surface and Interface Analysis* Published by John Wiley & Sons Ltd.

SURFACE and INTERFACE ANALYSIS

imaging,[31] and to ICR-MS spectrum data,[32] but we believe this is the first report of application of them to SIMS.

Time taken by RV1 compared to earlier algorithms

In Fig. 2 we compare calculation times for a relatively small but useful model problem of a 64×64 pixel image having an increasing size of spectrum associated with each pixel. This logarithmic plot indicates roughly power-law dependence of calculation time on the number of values in each spectrum. All calculations in this paper, including these, were performed on a Dell Precision T1700 desktop PC with 16GB of memory and an Intel Xeon E3-1271 CPU, running 64-bit Microsoft Windows 7. Matlab R2014a was used for most of the work, but we also used the portable version of Octave version 3.8.2. No conclusions should be drawn from this work about the relative speed of Matlab compared to Octave, because the 'portable' version of Octave does not give a fair comparison, instead it is the trend in calculation time as the spectrum size increases that is important, and this trend is very similar within both the Octave and Matlab platforms. Figure 2 does, however, show strong differences in speed between the available algorithms.

From Fig. 2 we can see that, especially for large spectra, the different algorithms available for calculating the SVD of the same matrix differ widely in their speed. In order of increasing speed for large matrices these are, (a) Octave xGESVD, (b) Octave xGESSD, (c) Matlab xGESVD, (d) Our implementation of Halko *et al.*'s algorithm (RV1) in Octave and (e) Our implementation of Halko *et al.*'s algorithm (RV1) in MATLAB. For images or datasets larger than those in Fig. 2, only Halko *et al.*'s algorithm is sufficiently fast for regular and routine use. In general waiting 20 s for a result (as in the longest calculation in Fig. 1(c)) is not a problem. Figure 1(e) offers the same result in about 0.1 s, a modest time saving. However, extrapolating Fig. 1(c) and Fig. 1(e) to larger problems—perhaps 512×512 pixels each with 60 000 mass values or more, is only really practical for the Halko *et al.* algorithm (which we have called RV1). Those of us that have tried to perform PCA on large datasets in the past, and found PCA to be impractically slow or demanding of memory can now reconsider. Halko *et al.*'s algorithms RV1 and RV2 remove these problems.

A numerical test-bed for applying new SVD methods to ToFSIMS images

We have developed our own code implementing both algorithms RV1 and RV2 as Matlab functions. We have integrated our code for RV1 as an option within the commercial MIA Toolbox and PLS Toolbox packages,[33] which provide an excellent graphical user interface for imaging multivariate analysis, and many tried-and-tested data pre-processing options. This new PCA option leads to a great reduction in calculation time in practical problems; however it is limited to cases in which the data is small enough to be held in memory, as required by RV1. By comparison RV2 has no such limit on the size of the data, but it would be difficult to take advantage of this 'out of core' version of Halko *et al.*'s algorithm without rewriting a large fraction of the pre-processing code in these toolboxes so as to avoid having to have the entire image data simultaneously in memory. Therefore we have taken the approach of verifying the analysis of smaller images (typically 256x256 pixels and below) in the MIA toolbox but we have written specialised MATLAB code to apply to larger problems using the RV2 'out of core' method. We shall now examine the results of applying RV1 and RV2 to examples of large ToFSIMS datasets.

Results of applying RV1: sodium citrate crystal

Figure 3(a) shows a total ion image for a small crystal of sodium citrate acquired using an Ionoptika J105 imaging ToFSIMS instrument (Ionoptika Ltd, Southampton, UK) and C60 primary ions at 40 keV. The crystal was mounted on double-sided tape, with some of this tape being visible on the left hand side of the image. Figure 3(b) is an optical microscope image of a similar (but not identical) sodium citrate crystal from the same sample. Figures 4(a) and 4(b) show positive ion SIMS spectra from the citrate crystal and tape respectively. A selection of results of applying RV1 to these images is shown in Fig. 5. In order to reduce the size of the data matrix to allow it to be held in memory for RV1 we have chosen to process a smaller region of interest (ROI) and narrower mass range than the originally recorded data contained, specifically a mass range from 100 to 200 amu. This is a mild form of 'peak picking'. This mass range seemed on inspection to contain the most useful information. No binning of mass values was done. The results show features not easily visible without PCA, such as a square region of damage visible within the 8^{th} and 10^{th} components because of an earlier rastered image in the same instrument. The 10^{th} component shows topographical features of the mounting tape on the left of the image.

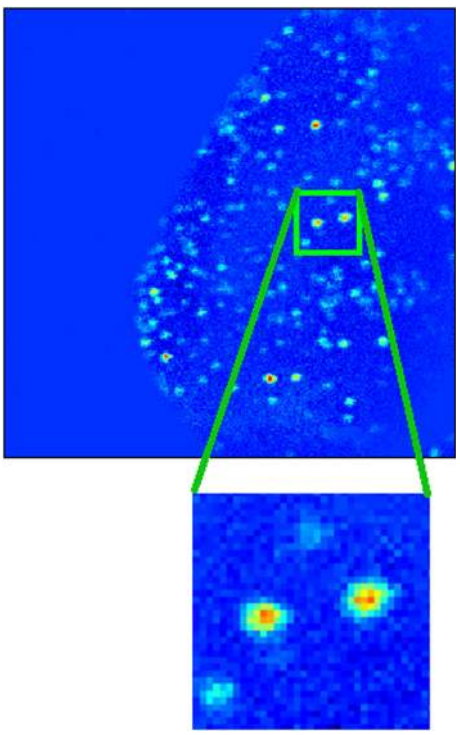


Figure 7. PCA analysis of the entire sodium citrate ToFSIMS image using the RV2 algorithm shows (in the 11^{th} component) small chemically distinct regions at the surface of the otherwise relatively homogeneous sodium citrate crystal. The field of view is 1 mm × 1 mm. The inset expanded region shows the enormous level of detail present in these RV2 data, making full use of the lateral resolution of the instrument.

© 2015 The Authors. *Surface and Interface Analysis* published by John Wiley & Sons Ltd.

Results of applying RV1: organic inclusions in basalt

Basalts from the Ontong Java Plateau possess easily distinguishable tube-like alteration textures that appear to be of biological in origin, but the unambiguous presence of molecular biomarkers or their digenetic products to date has yet to be fully confirmed. TofSIMS could be an extremely useful tool in the identification of these biomarkers. Figure 6 shows PCA results from applying algorithm RV1 to an edited area of an image acquired using 40-keV argon cluster ion as primaries in the Ionoptika J105, with a primary spot-size of around 3 μm. Fissures containing organic material are clear in components 1 and 2. Component 4, however, shows a purely instrumental effect unrelated to the chemistry of the surface. We believe this to be because of small temperature variations induced by the air-conditioning system causing expansion and contraction of the time-of-flight analyser. The J105 is not especially sensitive to such changes compared to other ToFSIMS instruments, but the laboratory was in the process of rearranging the air-conditioning system at the time of these measurements.

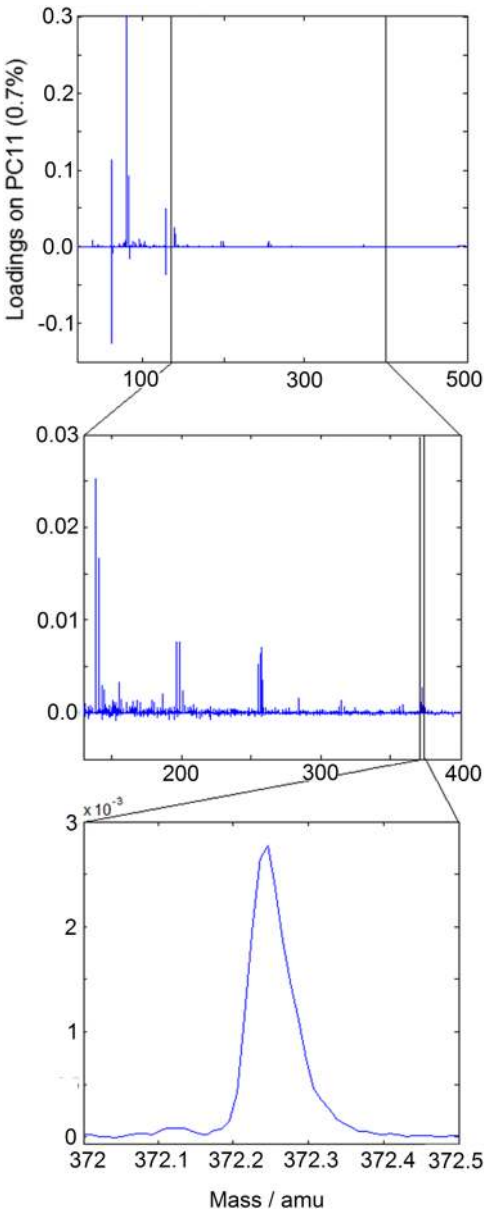


Figure 8. Basis spectrum (sometimes called 'Loadings') arising from the 11th component of the RV2 processed data corresponding to the image in Fig. 7. This (and the inset expanded regions) shows the full mass resolution of the instrument, with no binning needed.

Results of applying RV2

We applied RV2 to the entire basalt dataset of 512 × 512 pixels (a 2 × 2 tiled image of 256 × 256 pixels in each image) each pixel having a spectrum of 67 111 mass values. This took 6 h to process using RV2, and displayed similar results to those shown in Fig. 6 that took a few minutes to process using RV1. Applying RV2 to the sodium citrate image (taking about three hours) however produced more of a surprise. Figure 7 shows that the 11th component image indicates many small, chemically distinct features on the citrate crystal surface. Examination of the spectrum for this component, shown in Fig. 8, indicated that these small spots are likely to be NaCl, giving rise to peaks we attribute to $Na_3Cl_2^+$ ($m/Z = 138.91$), $Na_4Cl_3^+$ ($m/Z = 196.86$), $Na_5Cl_4^+$ ($m/Z = 256.82$) and $Na_7Cl_6^+$ ($m/Z = 372.24$). To show that the full mass resolution of the instrument is maintained by processing using RV2 we inset in Fig. 8 a high resolution expanded view around this last peak. The presence of these NaCl crystals on the surface of the citrate was a complete surprise and underlines the value of performing PCA on the whole data (using RV2) rather than editing the mass range and region of interest so as to reduce the size of the problem to fit in computer memory as required by previous methods.

Finally we should note that the design of the J105 instrument on which this work was performed is particularly suitable for multivariate analysis. In this design of ToFSIMS instrument the spectrometer and sample ionisation are separated so that the exact timing of secondary ion formation has no effect on the mass measured for that ion. Similar results should be possible on other types of instrument with careful calibration.

Conclusions

We have implemented, tested and verified two algorithms for PCA on large ToFSIMS (and other surface analysis) images. Algorithm RV1 is much faster than other algorithms, but may require some editing or binning of ToFSIMS data sets to enable them to fit in memory for processing. By comparison, RV2 is able to process ToFSIMS images of unlimited size (in fact limited only by the size of disc storage, typically several terrabytes in new desktop PCs). RV2 is necessarily slower, but still we find quite practical to include in the analysis of ToFSIMS images, because the longest processing time was still only around 6 h on a very ordinary desktop PC. These algorithms free us from the need to 'bin' mass spectra to lower resolution than they were acquired, and in the case of one sample (the sodium citrate crystal) showed the unexpected presence of NaCl crystals that would have been time-consuming and tedious to find by manually plotting the data in different ways.

We plan to make the MATLAB software described in this paper available from the NEXUS website (www.ncl.ac.uk/nexus) shortly after the SIMS XX conference this year. If any reader would like a copy of the software before then, please contact the corresponding author.

© 2015 The Authors. *Surface and Interface Analysis* Published by John Wiley & Sons Ltd.

Acknowledgements

The authors are very grateful to Dr John Fletcher for the SIMS images acquired from our samples using the Ionoptika J105 instrument in Gothenburg in the months before we took delivery of our own J105 at Newcastle. We thank Dr Graham Purvis of Newcastle University for the basalt rock sample. We thank users of the National EPSRC XPS Users' Service (NEXUS) for motivation in developing these methods for XPS applications, and EPSRC for funding the NEXUS Mid-Range Facility.

References

[1] R. Bro, A. K. Smilde, *Anal. Methods* **2014**, *6*, 2812–2831
[2] H. Tian, J. S. Fletcher, R. Thuret, A. Henderson, N. Papalopulu, J. C. Vickerman, N. P. Lockyer, Spatiotemporal lipid profiling during early embryo development of *Xenopus laevis* using dynamic ToF-SIMS imaging, *J. Lipid Res.* **2014**, *55*, 1970–1980.
[3] M. S. Wagner, D. G. Castner, Characterization of adsorbed protein films by time-of-flight secondary ion mass spectrometry with principal component analysis, *Langmuir* **2001**, *17*, 4649–4660.
[4] N. Sano, M.-L. Abel, J. F. Watts, The transfer of organics onto glass studied by ToF-SIMS, *Surf. Interface Anal.* **2011**, *43*, 423–426.
[5] K. Artyushkova, J. E. Fulghum, Angle resolved imaging of polymer blend systems: from images to a 3D volume of material morphology. *J. Electron. Spectrosc. Relat. Phenom.* **2005**, *149*, 51–60.
[6] D. Barbash, J. E. Fulghum, J. Yang, L. Felton, A novel imaging technique to investigate the influence of atomization air pressure on film–tablet interfacial thickness, *Drug Dev. Ind. Pharm.* **2009**, *35*, 480-486
[7] K. Artyushkova, J. E. Fulghum, Mathematical topographical correction of XPS images using multivariate statistical methods, *Surf. Interface Anal.* **2004**, *36*, 9.
[8] M. C. Biesinger, P.-Y. Paepegaey, N. S. McIntyre, R. R. Harbottle, N. O. Petersen, *Anal. Chem.* **2002**, *74*, 5711–5716
[9] A. Henderson, *Multivariate Analysis of SIMS Spectra in "TOF-SIMS: Materials Analysis by Mass Spectrometry"*, (2nd edition), (Eds: J.C. Vickerman and D. Briggs), SurfaceSpectra, Manchester and IM Publications, Chichester, **2013**, pp. 449–484
[10] D. J. Graham, D. G. Castner, Multivariate analysis of ToF-SIMS data from multicomponent systems: the why, when, and how. *Biointerphases*. **2012**; *7*(1): 49. doi:10.1007/s13758-012-0049-3.
[11] M. S. Wagner, D. J. Graham, B. D. Ratner, D. G. Castner, Maximizing information obtained from secondary ion mass spectra of organic thin films using multivariate analysis, *Surf. Sci.* **2004**, *570*, 78–97.
[12] J. Walton, N. Fairley, XPS spectromicroscopy: exploiting the relationship between images and spectra, *Surf. Interface Anal.* **2008**, *40*, 478-481.
[13] D. J. Graham, D. G. Castner, Image and spectral processing for ToF-SIMS analysis of biological materials. *Mass Spectrom*. **2013**; *2*(Spec Iss): S0014. doi:10.5702/massspectrometry.S0014.
[14] G. H. Golub, C. F. Van Loan, *Matrix Computations*, (3rd edn), Johns Hopkins, Baltimore USA, **1996**.
[15] J. Walton, N. Fairley, Noise reduction in X-ray photoelectron spectromicroscopy by a singular value decomposition sorting procedure. *J. Electron Spectrosc. Relat. Phenom.* **2005**, *148*, 29–40.
[16] A. J. Barlow, O. Scott, N. Sano, P. J. Cumpson, Multivariate Auger Feature Imaging (MAFI)—a new approach towards chemical state identification of novel carbons in XPS imaging, *Surf. Interface Anal.* **2015**, *47*, 173–175
[17] P. J. Cumpson, Angle-resolved XPS depth-profiling strategies, *Appl. Surf. Sci.* **1999**, *144*, 16–20
[18] P. J. Cumpson, Angle-resolved XPS and AES: depth-resolution limits and a general comparison of properties of depth-profile reconstruction methods, *J. Electron Spectrosc. Related Phenom.* **1995**, *73*, 25–52.
[19] J. Moore, Computational approaches for the interpretation of ToF-SIMS data. PhD Thesis, Manchester University UK, **2013**.
[20] M. R. Keenan, P. G. Kotula, Optimal scaling of TOF-SIMS spectrum-images prior to multivariate statistical analysis, *Appl. Surf. Sci.* **2004**, *231–232*, 240–244.
[21] D. J. Graham, M. S. Wagner, D. G. Castner, Information from complexity: challenges of TOF-SIMS data interpretation, *Appl. Surf. Sci.* **2006**, *252*, 6860–6868.
[22] B. T. Wickes, Y. Kim, D. G. Castner, *Surf. Interface Anal.* **2003**, *35*, 640–648.
[23] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, D. Sorensen, LAPACK users' guide, (3rd Edn), Society for Industrial and Applied Mathematics SIAM), **1999**.
[24] G. Golub, W. Kahan, Calculating the singular values and pseudo-inverse of a matrix, *SIAM J. Num. Anal. (Series B)*, **1965**, *2*, 205–221.
[25] G. H. Golub, C. Reinsch, Singular value decomposition and least squares solutions, *Numerische Mathematik* **1970**, *14*, 403–420.
[26] Ming Gu, Stanley C Eisenstat A divide-and-conquer algorithm for the bidiagonal SVD, *SIAM J. Matrix Anal. Appl.* **1995**, *16*, 79–92.
[27] N. P. Halko, P. G. Martinsson, J. A. Tropp, Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions, SIAM Rev., *Survey Rev. Sec.* **2011**, *53*, 217–288.
[28] N. P. Halko, Randomized methods for computing low-rank approximations of matrices, PhD Thesis, University of Colorado, **2012**.
[29] A. K. Cline, I. S. Dhillon, Computation of the singular value decomposition, Chapter 45 in *Handbook of Linear Algebra*, Edited by L. Hogben, Chapman & Hall/CRC: Boca Raton, USA, **2007**.
[30] Cleve Moler. 1976 Matrix Singular Value Decomposition Film. Los Alamos National Laboratory (Published on Youtube Dec 4, **2012**). https://www.youtube.com/watch?v=R9UoFyqJca8
[31] A. D. Palmer, J. Bunch, I. B. Styles, *Randomized Approx. Methods Efficient Compress. Anal. Hyperspectral Data Anal. Chem.* **2013**, *85*, 5078–5086.
[32] Lionel Chirona, Maria A. van Agthovenb, Bruno Kieffera, Christian Rolandob, Marc-André Delsuc, Efficient denoising algorithms for large experimental datasets and their applications in Fourier transform ion cyclotron resonance mass spectrometry, *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 1385–1390
[33] Eigenvector Research, Inc. 3905 West Eaglerock Drive, Wenatchee, WA 98801, USA, **2014**.

© 2015 The Authors. *Surface and Interface Analysis* published by John Wiley & Sons Ltd.