Cumpson PJ, Fletcher IW, Sano N, Barlow AJ. Rapid Multivariate Analysis of 3D ToF-SIMS Data: Graphical Processor Units (GPUs) and Low Discrepancy Subsampling for Large-Scale Principal Component Analysis. *Surface and Interface Analysis* 2016

DOI: [http://dx.doi.org/10.1002/sia.6042](http://dx.doi.org/10.1002/sia.6042)

Copyright:

© 2016 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](http://creativecommons.org/licenses/by/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Date deposited:

23/05/2016
Rapid multivariate analysis of 3D ToF-SIMS data: graphical processor units (GPUs) and low-discrepancy subsampling for large-scale principal component analysis

Peter J Cumpson,* Ian W Fletcher, Naoko Sano and Anders J Barlow

Principal component analysis (PCA) and other multivariate analysis methods have been used increasingly to analyse and understand depth-profiles in XPS, AES and SIMS. For large images or three-dimensional (3D) imaging depth-profiles, PCA has been difficult to apply until now simply because of the size of the matrices of data involved. In a recent paper, we described two algorithms, random vector 1 (RV1) and random vector 2 (RV2), that improve the speed of PCA and allow datasets of unlimited size, respectively. In this paper, we now apply the RV2 algorithm to perform PCA on full 3D time-of-flight SIMS data for the first time without subsampling. The dataset we process in this way is a 128 × 128 pixel depth-profile of 120 layers, each voxel having a 70 439 value mass spectrum associated with it. This forms over a terabyte of data when uncompressed and took 27 h to process using the RV2 algorithm using a conventional windows desktop personal computer (PC). While full PCA (e.g. using RV2) is to be preferred for final reports or publications, a much more rapid method is needed during analysis sessions to inform decisions on the next analytical step. We have therefore implemented the RV1 algorithm on a PC having a graphical processor unit (GPU) card containing 2880 individual processor cores. This increases the speed of calculation by a factor of around 4.1 compared with what is possible using a fast commercially available desktop PC having central processing units alone, and full PCA is performed in less than 7 s. The size of the dataset that can be processed in this way is limited by the size of the memory on the GPU card. This is typically sufficient for two-dimensional images but not 3D depth-profiles without sampling. We have therefore examined efficient sampling schemes that allow a good approximate solution to the PCA problem for large 3D datasets. We find that low-discrepancy series such as Sobol series sampling gives more rapid convergence than random sampling, and we recommend such methods for routine use. Using the GPU and low-discrepancy series together, we anticipate that any time-of-flight SIMS dataset, of whatever size, can be efficiently and accurately processed into PCA components in a maximum of around 10 s using a commercial PC with a widely available GPU card, although the longer RV2 approach is still to be preferred for the presentation of final results, such as in published papers. Copyright © 2016 The Authors Surface and Interface Analysis Published by John Wiley & Sons Ltd

Keywords: Time-of-Flight Secondary Ion Mass Spectrometry (ToFSIMS); Principal Component Analysis; PCA; Multivariate Analysis

Introduction

Principal component analysis (PCA) 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 peak positions, detecting outliers and finding patterns in massive datasets that are otherwise impossible to study by simply plotting one experimental variable against another. 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[9–11] 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.[12,13] Until now, these methods have been difficult to apply to very large datasets such as spectra associated with two-dimensional (2D) images or three-dimensional (3D) depth-profiles because the size of the dataset is too large to hold in the memory of commonly available personal computers (PCs). In recent work,[14] we applied the new ‘random vectors’ (RVs) method of SVD proposed by Halko[15] and co-authors to time-of-flight (ToF)-SIMS images for the first time. One variant of this ‘RVs’ algorithm, which we called RV1, increases the speed of calculation by a factor of several hundred, making PCA of these datasets practical on desktop PCs for the first time. More important, a second variant of this algorithm, RV2, allows any size of dataset to be processed by an ‘out-of-core’ method. When applying RV2, the ToF-SIMS data are stored on disc and brought into memory piece-by-piece for processing. This ensures that there is no limit (except the disc capacity itself) on the size of the dataset that can be processed. The ‘out of memory’ errors common in the past have been eliminated. This is especially useful when, for example,
preparing data for publication or in a report, when one can wait a few hours or even a few days for RV2 to perform a complete and rigorous PCA analysis of the data. However, there are occasions where time is much more limited. Therefore, one purpose of this paper is to investigate how to do SVD (and therefore PCA) using the RV1 algorithm as rapidly as possible using modern commercially available PC hardware. We do this by

1. applying graphical processor units (GPUs) not previously used in surface analysis to increase the speed of PCA calculation by a factor of around 4.1 compared with the same PC without such a GPU card, and
2. for the largest datasets, we apply a new sampling method to select a subset of pixels (or voxels) for processing that converges more rapidly than previously used sampling methods.

Methods (a) and (b) used individually or together allow us to apply PCA to moderate-sized ToF-SIMS datasets in a reasonable time (up to about 10 s) on a high-specification PC with a GPU card. For PCA analysis of a very large datasets where the longer calculation time is not an issue, we would still recommend using the RV2 algorithm described previously. RV2 runs well on any specification of PC, even laptops, and does not use any GPU card, but can take hours to complete processing a dataset because of the time taken fetching segments of the dataset from disc.

Requirements of surface analysis

We need a faster SVD algorithm for large images and especially 3D datasets. We can therefore identify three key requirements of an algorithm suitable for 3D imaging applications:

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

Note that there is an extensive literature and recommendations on data pre-processing required in XPS and ToF-SIMS, which may include normalisation, mean-centering and Poisson distribution variance correction. 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 and therefore pre-processing will not be considered here.

A 3D ToF-SIMS test dataset

As a set of test data, we chose a 20 kV C$_{60}$ imaging depth-profile of a plant leaf using our Ionoptika J105 chemical imaging ToF-SIMS instrument in positive ion mode. The primary beam current was 5 pA, and the repetition rate was one pulse every 100 μs, with no sample bias. This dataset has 128 x 128 pixels in x and y, each pixel being 2 μm x 2μm, 120 levels in z, and a mass spectrum of 70 439 values at each voxel, with this mass range covering the interval 30 to 600 amu.

Assuming 8 bytes per mass spectrum value (as would be required for double-precision representation by most types of pre-processing), this means a dataset over 1 terabyte in size (although in most cases ToF-SIMS software will save this in a numerically compressed form).

Figure 1 shows some key features that can be seen by plotting the intensity of particular chosen peaks from within this enormous dataset. In particular, there appears to be a hydrocarbon-rich (and therefore perhaps ‘waxy’) cuticle at the leaf surface (indicated by, for example, C$_2$H$_4$ at m/z 91) above a region that gives rise to large range of Ca-containing ions and complex ions (e.g. Ca$_2$O$^-$) that appear to indicate the location of cytoplasm within cells below the surface. Notice the dark horizontal band in the y–z plot shown in Fig. 1. This is an instrumental artefact caused by a temporary decrease in the primary beam current. We will not attempt to remove this by any normalisation or other correction, so the same band can be seen (and easily discounted) in the presentation of PCA results below.

We processed this entire dataset (not just the small subset of that data shown in Fig. 1) including all of the mass spectra in full, in 27 h on an ordinary Windows PC (a Dell Optiplex 7010 having 16 Gb of memory) using the RV2 algorithm previously described. We calculated the first 100 principal components. When using RV2, the speed of processing is not very dependent on the specification of the central processing unit (CPU), but more dependent on the speed of reading from disc, so that a solid state disc drive offers a speed advantage, although in our case, the leaf data was processed using a conventional hard drive.

It is important to note that the specific mode of operation of the J105 instrument makes PCA rather easier to apply than in some other designs of ToF-SIMS instrument. The J105 performs decoupling of the timing of the primary beam from the timing of ion pulses entering the mass analyser, avoiding calibration problems due to spatial distribution of formation of secondary ions and allowing the use of PCA with no need for prior calibration specific to the sample (although of course the mass scale of the instrument itself is calibrated). In other instrument designs, where the arrival time of ions in the analyser with respect to the impact of the primary ions on the sample is crucial, additional careful calibration may be needed to ensure valid PCA results.

Figures 2 to 8 show PCA results, including orthogonal slices in x, y and z through the PCA component scores obtained from this calculation. PC1 represents the average of all of the spectra and was dominated by the relatively intense signals from the inorganic species present in the sample. These were mainly due to calcium species and arose from beneath the leaf surface. PC3 and PC5 (as well as some higher components not shown) are dominated by signals due to various Ca species. These presented as several series of cluster ions formed in the SIMS ionisation process containing Ca, O and H. Ions of the general form [(CaO)$_x$(H$_n$H$_{12}$O)$_y$]$^+$ and [Ca (CaO)$_x$(H$_n$H$_{12}$O)$_y$]$^+$ were detected with $x = 1$ to 9 and $n = 0$ up to 8. Differences in the relative intensity distributions of these species may well reflect differences in the Ca chemistry within the leaf structure. PCs 2 and 4 were dominated by relatively weak signals from various organic and hydrocarbon species. Consideration of Figs 4 and 7 shows that the majority of these signals arose from the leaf surface region – this would be consistent with the presence of a waxy cuticle. Other signals in PC2 arose from the cell walls that would be consistent with a cellulosic material or similar. The remaining signals in PC4 were due to Ca species and appeared to be associated with the cell contents.

Figure 5 shows loadings from PC2, plotted in blue for the positive part of the loading and red for the negative part. Even this is only a section from the entire 1 to 600 amu mass range, with only 70 to 160 amu being shown in this figure, nevertheless we can see the high mass-resolution of the mass spectral dataset at m/Δm ≈ 1500. Also the blue positive part of PC2 is dominated by organic fragment ions at higher than the nominal mass while the red negative part is
dominated by inorganic cluster ions at lower than the nominal mass (i.e. \( \text{Ca}_2\text{O}_2^+ \) and \( \text{Ca}_2\text{O}_2\text{H}^+ \) at \( m/z \) 112 and 113, respectively). There was also an instrument tuning issue that gave rise to split peaks/poor peak shapes in the mass spectra, but this had no effect on the results presented in this paper. This effect can be seen in Fig. 5 especially for the fragment ion peaks at \( m/z \) 109 and 115 due to \( \text{C}_8\text{H}_13\text{O}_{13}^+ \) and \( \text{C}_9\text{H}_7\text{O}_{27}^+ \), respectively. At the time these data were acquired, the instrument was very new and had a minor internal charging issue due to some surface contamination on some internal surfaces of the in the ToF. This has now been fully resolved with thorough cleaning and reassembly, but we do not expect this to have affected the performance of PCA on these data.

Figures 2 to 8 show the results of applying PCA to a dataset around 20 times larger than any that we know of as having previously appeared in the literature.
A dataset that can be processed using the RV2 algorithm is the capacity of the hard disc. Faster, if approximate, PCA results would be useful in many cases – therefore in the remainder of this paper, we will look first at accelerating the calculation using a GPU. Then, for the largest datasets, we look at efficient sampling to reduce the time taken in calculation.

Figure 6 the third principal component seems (as with the first) to be located with the contents of plant cells in the leaf, possibly showing slightly higher spatial resolution. Blue areas show where this score is positive, and red where it is negative. Both the positive and negative loadings of PC3 contain Ca species.

**Time taken for the RV1 algorithms using CPU and GPU**

In Fig. 9, we compare calculation times for a very typical ToF-SIMS multivariate analysis problem of a 256 × 256 pixel image having an increasing size of spectrum associated with each pixel. This is a large PCA problem but is still small enough (unlike our leaf data earlier) to be processed in memory by both CPU and GPU processors using the RV1 algorithm, given their memory limitations. All calculations in this section were performed on a Hewlett Packard Z440 desktop PC with 132 GB of memory and an Intel Xeon E5-1620 v3 CPU having eight cores operating at 3.5 GHz, running 64-bit Microsoft Windows 7. MATLAB R2014b was used. The GPU card is an Nvidia Tesla K40 with 2880 processors and 12 Gb of memory.

Clearly, both the CPU and GPU methods involve significant parallel processing, and monitors of CPU and GPU activity showed a fairly even spread of the workload across them. Overall, the use of the GPU accelerates the calculation by around a factor of 4.1 compared with the CPU alone. Thus, a very typical dataset, 256 × 256 pixels each with 20,000 mass values, can be processed in around 7 s using the GPU card, including the time taken to move the large volume of data from PC to GPU card and retrieve the results. This is very much...
and the use of virtual memory leads to prohibitive calculation times.

The need for sampling

Unless you need to apply PCA to the largest datasets, you may not need to read this section. One obvious method of reducing the computational workload involved in processing the largest datasets is to sample only a proportion of the pixels (or voxels) in the data, as shown schematically in Fig. 10. If we perform PCA on this sample (not the whole dataset), we can then apply the orthogonal basis set so obtained to the entire data. This is sometimes called using a ‘training set’, for example, Van Nuffel et al \cite{69} used a training set of 6.1% of a 3D ToF-SIMS dataset, with voxels being selected from the full dataset at random. The entire spectrum from each sampled voxel (or pixel in the 2D case) is included as a row in the data matrix, but only a small number of voxels are sampled, and therefore the matrix is much smaller than if all the voxels were to appear as rows. The full dataset could be presented in terms of the projection of the spectrum in each pixel (or voxel) onto the principal component loadings. The result, ideally, would be very little different to seeing the PCA results of the entire data. Provided the sampled voxels (or pixels in the 2D case) are representative of the data as a whole, and there are more samples than the rank of the full data matrix (preferably many more), then, we should expect that the basis calculated by PCA will represent the data as a whole too, and nothing will be lost by using a sample of the data rather than the data as a whole. This is a risky assumption to make without knowledge of the source of the imaging data; however, in the specific case of ToF-SIMS imaging, we know that there is a fundamental diameter of the primary beam and sputtered volume limit below that pixels will be highly correlated. In the remainder of this paper, we will apply sampling schemes to our ToF-SIMS leaf dataset to examine the advantages and limitations of sampling, as well as which is the best method of sampling, as we shall now discuss.

Qualitative discussion of sampling methods

Key to the success of sampling is the selection of a suitable sampling method to choose the voxels (or pixels) to use. It is possible to sample using a regular grid. For example, one could include every kth voxel in the PCA calculation. This can be a good method, but requires careful choice of the grid to apply (i.e. a careful choice of k) in advance, and unfortunately one does not know the spatial extent of the components of the image in advance. For example, consider the 2D sampling grid shown in Fig. 11. In this case, we have sampled 500 pixels from an image of 256 × 256 pixels at regular intervals of (256 × 256)/500 or around k = 131 pixels. Each sampled pixel is indicated by a small circle in Fig. 11. Clearly, this forms a pattern of repeated grid lines.

It would be good to make sampling more homogenous and less directional than this grid method. The most obvious way to do this is to sample pixels (or voxels in 3D) at random. Figure 12 shows a random selection of 500 pixels from the same set of 256 × 256 shown in Fig. 11.

This time, there is no preferred direction to the acquired data, even on the fastest ToF-SIMS instruments commercially available. Importantly for GPU acceleration, our PCA calculation applies the same instructions simultaneously to many segments of data, making it very suitable for full utilisation of these processors. If many different instructions needed to be executed simultaneously, it would most likely not be an advantage to use a GPU card.

Figure 8. The fifth principal component, PCS, scores.

Figure 9. Calculation times recorded for principal component analysis of complete 256 × 256 pixel images having increasing spectra associated with each pixel. The RV1 algorithm previously described to calculate 100 principal components using the central processing unit (CPU) and graphical processor unit (GPU) of the personal computer. For spectrum lengths typical of ToF-SIMS images (around n = 2000 to 50 000), the GPU performs principal component analysis remarkably quickly, but beyond the capacity of the GPU memory, the CPU still performs well. For the largest problems (n > 100 000), the 128 Gb memory of the personal computer becomes full, and the use of virtual memory leads to prohibitive calculation times.
where species ‘A’ can reside within the image without being sampled at all, whereas region ‘B’ is sampled excessively and is overrepresented in the subsequent PCA analysis.

The sampling problem here has features similar to those of multidimensional Monte Carlo integration, where methods based on low discrepancy series (LDS) have proved increasingly useful in recent years. LDS can be viewed as quasi-random number series that have poor randomness properties—they are in a sense quite poor random numbers. There are several possible LDS that may be used in our problem of sampling prior to PCA, but we have most often used one based on Sobol series.\[20,21\] We have previously used Sobol series in high-dimensional integration problems that arise in Monte Carlo integration of simulated electron trajectories in XPS.\[22\] Figure 14 shows pixels sampled from the same 256 × 256 by a Sobol sequence.

One can see that, speaking roughly, the pixels sampled by this Sobol series fill the area more evenly than the random sampling scheme does. Formally, the discrepancy of a series is computed by comparing the actual number of sample points in a given volume of multidimensional space with the number of sample points that should be there in a long-term limit of a uniform distribution. We would expect that the use of an LDS results in more rapid convergence towards the true singular values, eigenvalues and eigenvectors describing the data than would random sampling, and this is indeed what we observe in analysing the ToF-SIMS data. Sobol (and other LDSs) are easily extended to 3D.

Low-discrepancy sequences originated in the context of numerical integration of functions in high-dimensional spaces.\[23\] Quasi-Monte Carlo integration uses an LDS such as the Halton\[24\] sequence or the Sobol sequence whereas conventional Monte Carlo uses a pseudorandom sequence. The function is sampled at points within the limits of integration according to these sequences, and the values of the function summed numerically. The advantage of using LDSs is a faster rate of convergence in this integration application, often proportional to \(N^{-1}\) in the limit of large sample size \(N\) (as would be the case for sampling on a regular grid), whereas for random sampling, this is only around \(N^{-0.5}\). We conjecture that similar convergence behaviour will be seen in applying the low-discrepancy versus random sampling methods in...
SVD of large spectral imaging datasets for PCA. For the large sample sizes necessary in the PCA of ToF-SIMS datasets, this difference is very important. Since originally proposing this combination of PCA and LDS at the SIMS XX conference in Seattle in 2015, we have found previous published work that indicated that this combination has worked well in the entirely different context of quantitative financial analysis\cite{25} and Monte Carlo methods\cite{26}.

Comparing convergence of Sobol vs. random sequences

We will compare the application of random versus LDS methods to our 3D leaf ToF-SIMS dataset. Because of the more rapid convergence of LDS sampling in numerical integration applications, we would expect more rapid approximation of the PCA components of a 2D or 3D dataset when an LDS sampling method is used. This is the first time that we know of that LDS methods have been applied in surface analysis.

Figure 15 shows the locations within the 3D dataset of 20,000 voxels sampled from the total of $128 \times 128 \times 120 = 1.96$ million voxels in the leaf 3D dataset.

As discussed earlier, SVD is at the core of PCA calculation, whereby we can express a matrix as

$$A = USV^T$$

where $A$ is the ‘design matrix’ containing the spectra, $U$ 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. In practical application of PCA to surface analysis, typically there are between roughly two and ten important principal components, although sometimes more. Therefore, a useful metric to assess how accurately PCA of a sampled dataset approximates the PCA components of the full dataset may be the following: we add the root-mean-square errors in the ratio of the second, third, … tenth singular values ratioed to the first, i.e.

$$W = \sqrt{\sum_{i=2,3,\ldots,10} \left( \frac{S_{\text{sampled}}(i,i)}{S_{\text{sampled}}(1,1)} / \frac{S_{\text{full}}(i,i)}{S_{\text{full}}(1,1)} \right)^2}$$

where $S_{\text{full}}$ and $S_{\text{sampled}}$ are the singular value matrices from SVD of the full and sampled design matrices, respectively. If the sampling produces a good approximation to the singular values (and eigenvalues) of the full problem, then $W$ is small. We propose this performance metric as an indicator of how closely the sampling method reproduces the true ratios of the quantity of the different chemical species that show up as components in the PCA analysis. Other equally good metrics are no doubt possible, but this seems a reasonable one. We evaluated $W$ numerically for 500 separate PCA calculations on the leaf 3D dataset. In each case, the full mass spectra (each voxel having 70,439 values) were used, but the number of voxels sampled was increased from 0.01% to 1% of the full dataset in 500 steps. This allowed all calculations to be performed rapidly using the GPU card, as 1% of the dataset represents just under the 12 Gb memory capacity of our GPU card, allowing the RV1 algorithm to be used. To evaluate $W$, we used $S_{\text{full}}$ taken from our earlier (and much more time-consuming) RV2 calculation from the full 3D dataset of over 1 terabyte in size. In each case, the full mass spectra (each voxel having 70,439 values) were used, but the number of voxels sampled was increased from 0.01% to 1% of the full dataset in 500 steps. This allowed all calculations to be performed rapidly using the GPU card, as 1% of the dataset represents just under the 12 Gb memory capacity of our GPU card, allowing the RV1 algorithm to be used. To evaluate $W$, we used $S_{\text{full}}$ taken from our earlier (and much more time-consuming) RV2 calculation from the full 3D dataset of over 1 terabyte in size. Figure 16 shows how $W$ decreases gradually as we increase the number of sampled voxels for both the random sampling and Sobol sampling methods. In the higher part of this range, the Sobol sampling method leads to a smaller error than random sampling on average, although there is naturally some random variability in both. Figure 17 shows the ratio of $W$ for Sobol compared with random sampling. The most
important regime is the larger sample size range here, and taking an average over the range 0.5 to 1.0% sampling, we find that Sobol sampling gives about 0.58 of the root-mean-square error we observe from random sampling, corresponding effectively to an acquisition time reduced by a factor of about $(1/0.58)^2 \approx 3$. Clearly, the exact improvement offered by Sobol sampling will be dependent on many parameters, but is likely to be an advantage in most practical studies.

**Comments on slow, ultra-high mass resolution SIMS spectromicroscopy, and the role of LDS sampling**

While the ToF analyser is now common in SIMS, there are some options in the future for much higher mass resolution based on Fourier-transform mass spectrometry (FTMS) analysers. There are few demonstrations of successful integration of FTMS analysers at present, with ion cyclotron resonance mass spectrometry (ICR-MS) being the only published example\cite{27,28} that we know of at the time of writing; however, there is talk of integrating quadro-logarithmic electrostatic traps and other types of ion-trap, and a commercial option has recently become available.\cite{29} Although not widely available as yet, it is possible (although by no means certain given the commercial and cost issues) that these FTMS analysers will become popular additions to ToF-SIMS instruments in the future. These analysers offer extremely high mass resolution at the expense of longer acquisition times than the ToF analyser. These longer acquisition times can often be in the range 100 ms–10 s per analytical point, making image acquisition with these types of analyser very slow for, say, 256 × 256 pixel images and prohibitive for full 3D datasets. The previous discussion on the application of LDS sampling has some interesting implications for the use of these slow, high resolution analysers in conjunction with ToF-SIMS imaging. If one has a sample that one wishes to image using both ToF (fast) and FTMS (slow) analysers, and one begins with no initial knowledge of the sample, then it may be useful to use a Sobol (or other LDS) sequence to select pixels (or voxels) to be analysed by FTMS analysis, while the whole image (or 3D dataset) is acquired for ToF analysis. LDS sampling in this way has some nice properties such as avoiding sampling pixels in close proximity, while uniformly covering the imaged area. This application is reminiscent of the biological problem of placement of retinal photoreceptors by a ‘blue noise’ distribution that avoids having
pairs of photoreceptors too close and yet allows high spatial frequencies to be detected without aliasing.\textsuperscript{[20]}

Taken together, data from both ToF and FTMS analysers may be processed very efficiently by the PCA methods we have discussed to make best use of the speed of the ToF and the mass resolution of the FTMS together. If, on the other hand, the ToF-SIMS imaging is done first and PCA performed on that image, then the location of regions rich in the chemical components revealed by PCA may be suitable choices of location for FTMS analysis. In that case, one is using prior knowledge of the sample gained from ToF-SIMS to guide the FTMS sampling.

**Conclusions**

We have demonstrated PCA on a large 3D ToF-SIMS dataset, and how GPU and LDS techniques may be used to obtain a rapid and accurate approximation to the PCA results from the full dataset. The most rapid PCA is performed using a PC having a GPU card executing algorithm RV1. This GPU card must have enough memory to accommodate the full, uncompressed dataset. A typical dataset of $256 \times 256$ pixels and 2000 mass values can be accommodated on the GPU card we used, with 12 Gb of local memory. It is therefore important to ensure that any GPU card one uses has sufficient memory as well as processor cores.

For datasets larger than the GPU card can accommodate, modern PCs with large physical memory work well, again using RV1, but are about four times slower. If the dataset is too large to accommodate in the physical memory of the PC, then one has two options:

1. Process the entire dataset using algorithm RV2. This will often take several hours, sometimes up to a day or two for the largest datasets. The speed of processing is less dependent on the specification of the CPU, and more dependent on the speed of reading from disc, so that a solid state disc drive offers a speed advantage: or

2. Subsample, ideally using a low-discrepancy method such as Sobol sampling, making a dataset sufficiently small to fit in the physical memory of the PC or GPU card, and then use RV1. This will typically reduce the time taken to between 5 and 10 s, so is particularly useful in the middle of taking spectra or images to inform decisions about where and how to take the next data.

Finally, we should make one observation about these techniques and their take-up by the biological and medical communities. The use of high spatial and mass-resolution IMS instruments together with multivariate analysis (see Figs 2 to 8 for example) offers a new and powerful tool for looking at the biochemistry of slices through 3D depth-profiles. This is a technology not well appreciated yet in biomedical communities, and therefore in an attempt to more precisely convey their power, we have presented these types of results under the name ‘tomomics’ internally at our university. Tomomics combines the Greek τομ(ο)ς (tomas) meaning ‘slice’ or ‘section’ (from which we also derive the word ‘tomography’) with an ‘-omics’ suffix. The Oxford English Dictionary defines nouns having this suffix in the context of cellular and molecular biology as having the sense of ‘all constituents considered collectively’. This seems very appropriate to the multivariate analysis of all the information in 3D IMS datasets. In speaking to communities who are primarily interested in what results the technique is capable of rather than the details of instrumentation or software, and who are comfortable with the meaning of genomics, proteomics and other ‘-omics’ fields, the term ‘tomomics’ may be useful.

**Acknowledgements**

Preliminary X-ray photoelectron spectroscopy (XPS) data were acquired at the National EPSRC XPS Users’ Service, an EPSRC Mid-Range Facility. The authors are very grateful to Paul Blenkinsopp of Ionoptika Ltd for assistance in acquiring the ToF-SIMS datasets using our Ionoptika J105 instrument now at Newcastle. This ToF-SIMS instrument was purchased as part of an instrument package funded by EPSRC’s ‘Great Eight’ capital funding grant EP/K022679/1 and Newcastle University.

**References**