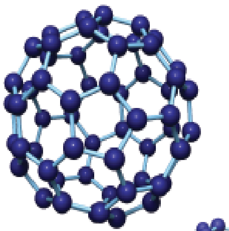
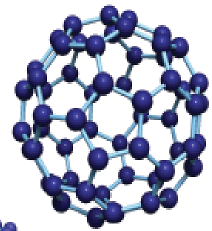
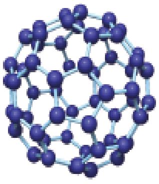
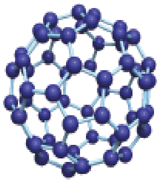
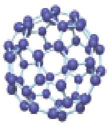
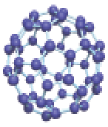
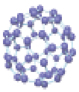
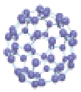
Computational Approaches for the Interpretation of ToF-SIMS Data.

JIMMY MOORE

SUPERVISORS:



DR. ALEX HENDERSON AND DR. NICK LOCKYER,

SURFACE ANALYSIS RESEARCH CENTRE (SARC).

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN THE DEPARTMENT OF CHEMICAL ENGINEERING AND ANALYTICAL SECIENCE (CEAS),

UNIVERSITY OF MANCEHSTER.

SUBMITTED: 2013

Contents

word count = 40,577

List of Figures

List of Tables

List of Abbreviations

Analogue to Digital Converter	ADC
Atomic Mass Unit	u
Canonical Variates Analysis	CVA
Chemical Ionization	CI
Continuous Wavelet Transform	CWT
Cross Validation	CV
Dalton	Da
Discriminant Function Analysis	DFA
Electron Ionization	EI
Electrospray Ionization	ESI
Fast Atom Bombardment	FAB
Field Ionization	FI
Fourier Transform Ion Cyclotron Resonance Mass Spectrometry	FTICR-MS
Gas Chromatography	GC
Graphical User Interface	GUI
Leave-One-Out-Cross-Validation	LOOCV
Liquid Chromatography	LC
Liquid Chromatography-Mass spectrometry	LC-MS
Mass spectrometry	MS
mass-to-charge ratio	m/z

Matrix Assisted Laser Desorption and Ionization	MALDI
Multivariate Analysis	MVA
Multivariate Curve Resolution	MCR
Neural Networks	NNs
Partial Least Squares	PLS
Photon Ionization	PI
Principal Component Analysis	PCA
Principal Component Discriminant Function Analysis	PC-DFA
Principal Components	PCs
Random Access Memory	RAM
Secondary Ion	SI
Secondary Ion Mass Spectrometry	SIMS
Time of Flight Mass Spectrometry	ToF-MS
Time to Digital Converter	TDC
Undecimated Discrete Wavelet Transform	UDWT

Abstract:

High surface sensitivity and lateral resolution imaging make Time-of-Flight Secondary Ion Mass Spectrometry (ToF-SIMS) a unique and powerful tool for biological analysis. Many of these biological systems, including drug-cell interactions, require both the identification and location of specific chemicals. ToF-SIMS, used in imaging mode, is making great strides towards the goal of single cell and tissue analysis. The experiments, however, result in huge volumes of data. Here advanced computational approaches employing sophisticated techniques to convert these data into knowledge are introduced.

This thesis aims to produce a framework for data analysis, integrating novel algorithms, image analysis and 3D visualisation. New schema outlined in this thesis address the issues of the immense size of 3D image stacks and the complexity contained within the enormous wealth of information in ToF-SIMS data.

To deal with the issues of size and complexity of ToF-SIMS data, new techniques to processing image data are investigated. Automated compression routines for ToF-SIMS images using a peak picking routine tailored for ToF-SIMS are evaluated. New user friendly GUIs capable of processing and visualising very large image stacks are introduced as part of a tool-kit designed to streamline the process of multivariate analysis and image processing. Along with this two well known classification routines, namely AdaBoost and SVMs, are also applied to ToF-SIMS data of several bacterial strains to test their ability to classify SIMS data accurately. This thesis present several new approaches to data processing and interpretation of ToF-SIMS data.

Declaration:

Computational Approaches for the Interpretation of ToF-SIMS Data.

Supervisors:

Dr. Alex Henderson and Dr. Nick Lockyer

This thesis is presented in partial fulfillment of the requirements for the completion of a PhD.

“No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning;”

Submitted: 2013

COPYRIGHT STATEMENT:

Publications List:

- Jimmy D. Moore, Alex Henderson, John S. Fletcher, Nicholas P. Lockyer, and John C. Vickerman. Peak picking as a pre-processing technique for imaging time of flight secondary ion mass spectrometry. Surface and Interface Analysis, 45(1):461:465, 2013.

- Alex Henderson, Jimmy D. Moore, and John C. Vickerman. SIMS informatics. Surface and Interface Analysis, 45(1):471:474, 2013.

- Edward G. Chadwick, N. V. V. Mogili, Colm O'Dwyer, Jimmy D. Moore, John S. Fletcher, Fathima Laffir, Gordon Armstrong, and David A. Tanner. Compositional characterisation of metallurgical grade silicon and porous silicon nanosponge particles. RSC Adv., 2013.

Acknowledgements:

I would like to thank the EPSRC and Prof. John Vickerman for my funding.

A huge “thank you” to everyone in the group for a warm welcome and great working atmosphere.

Special thanks to Nick for his support throughout my time with the group.

Thanks to John Fletcher for his endless knowledge, helpful talks and his sweet guitar playing.

Thanks to Paul Bassan for his insightful chats and general helpful input.

Massive thanks to Alex for his thought provoking talks, his vast understanding of all things multivariate and for his continued understanding and support in and out of the office.

Finally, a big thank you to Stephen for everything.

Chapter 1

Introduction

Contents

1.1 Mass Spectrometry

Mass spectrometry (MS) is an analytical technique that is used to determine the mass of chemical compounds. Since the mass of a chemical compound is dependent on its elemental composition, it can be an important measure for its identity. Since the advent of mass spectrometry in the early 1900's [71] it has gradually grown from a very specialist analytical technique into a mature method that is widely used and one of the most powerful tools for the identification of unknown compounds. Fields of application include biology, medicine, forensics, organic and inorganic chemistry, materials engineering, food engineering and others. MS is a widely used tool, both in academia and in industry.

MS has the advantage that it does not necessarily have to be used as a stand-alone technique, but can be combined with one or several steps of separation techniques. These separation techniques include gas chromatography (GC), liquid chromatography (LC), solid phase microextraction and virtually any other analytical separation technique. Coupling of separation techniques to MS can either be indirect, in which separation is done in a stand-alone experiment and followed by a MS experiment, or directly in which both are combined in a single setup. The latter method (e.g. LC-MS) is very common for routine analysis of diverse complex mixtures.

To determine their mass, chemical compounds need to be brought into the gas phase, ionized, separated based on their mass (mass-to-charge ratio) and detected. All these steps are crucial in the successful generation of a mass spectrum. Gas phase ions are typically generated in vacuum or transferred into vacuum directly after they are created. Methods to do this include, but are not limited to, Secondary Ion Mass Spectrometry (SIMS), Electron Ionization (EI), Electrospray Ionization (ESI) and Matrix Assisted Laser Desorption and Ionization (MALDI).

The combination of several ionization techniques can be used to combine their specific advantages and generate complementary data. In most cases, the analytes are brought into the gas phase and ionized in a single step. After bringing ions into the gas phase and into the vacuum, they have to be separated based on mass and detected. Mass analysis can be done using sequential mass filtering, in which ions that fit in a single mass window are transmitted and all other ions are blocked. The mass window is scanned and ions are detected for each mass window. Sequential filtering methods include quadrupole filtering and magnetic sector filtering. More efficiently, all present gas phase ions are analysed in a single MS-run. The latter method is much more economic because the loss of ions that is inherent to the sequential filtering method is circumvented. Methods that detect all ions in a single measurement step include Time of Flight Mass Spectrometry (ToF-MS), Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FTICR-MS) and Orbitrap mass spectrometry. Combinations of mass filters and mass analysers are also quite common.

All mass spectrometers measure mass-to-charge ratio (m/z) as opposed to simple mass. This IUPAC standard m/z denotes the quantity formed by dividing the mass of an ion by the unified atomic mass unit (u) and by its charge number (positive absolute value)[73]. Masses tend to be singly charged, however this is not always the case. If an ion is more highly charged for example if z (charge) = 2, then that ion will appear at half its mass. The larger or more intense a peak is in the spectrum, the greater the abundance of that ion. Mass spectrometry uses the atomic mass unit as the standard mass unit, though the Dalton (Da) is also used.

This form of analysis produces a mass spectrum of the constituents within the sample i.e. the mass spectrum generally contains data corresponding to the masses of fragmentation products of a molecule. The nature of fragmentation process relates quite specifically to the structure of the molecule and the technique used. This chemistry is then used to deduce the structure of the unfragmented molecule or parent material.

In brief summation: measurement of a mass spectrum requires ionization of a sample, filtering the reactant and product ions based on mass and charge, detection of the ions after filtering and storing of the data. Many techniques are used to ionize molecules for mass spectrometry. These include Secondary Ion (SI), Electron

Ionization (EI), Photon Ionization (PI) , Field Ionization (FI) , Chemical Ionization (CI) , Electrospray ionization (ESI), Fast Atom Bombardment (FAB) and others. Understanding the ionization process is crucial to understanding the mechanisms by which mass spectral data are produced. Once formed, ions are accelerated and focused into a mass analyser. There are several common ways to accomplish mass analysis: time-of-flight, magnetic field, quadrupole field and ion cyclotron resonance. This thesis focuses on SIMS, specifically ToF-SIMS but most of the content is applicable to other forms of MS.

1.2 Time-of-Flight Secondary Ion Mass Spectrometry

ToF-SIMS is the mass spectrometry of ionised particles that are emitted when a surface is bombarded by energetic primary particles/ions. The secondary ions are removed from layers on the surface of a sample. These ions are then accelerated into a “flight tube” and their m/z are determined by measuring the time at which they reach the detector (i.e. Time-of-Flight). A brief explanation of the technique is given in this section. A thorough explanation of the TOF-SIMS technique is given in a book edited by Vickerman and Briggs [78].

SIMS is a widely used technique to analyse the chemical composition of solid surfaces and thin films. It can be applied to analyse both the elemental and the molecular composition of the surface. The emitted (sputtered) secondary particles are electrons; neutrals species atoms or molecules; atomic or cluster ions. The vast majority of sputtered particles are neutrals, but it is the secondary ions that are analysed and detected by a mass spectrometer. Ion beam induced chemical damage to the surface plays an important role. This has had a large effect on the history of SIMS owing to the fact of damage arising due to primary ion choice. Thus, Static SIMS, a regime of SIMS developed in the late 1960s in Benninghoven’s group [4] in Münster, in which only a small part ($< 1\%$) of the surface is exposed to the primary ion beam, was introduced. In this regime, each measured secondary ion originates from an untouched part of the surface and is therefore representative of the virgin surface. The possibility to do static SIMS, extended the SIMS technique from an elemental analysis technique to a molecular analysis technique. With the introduction of cluster ion beams the capability of SIMS instruments has increased dramatically to allow analysis beyond the static regime [5] [31].

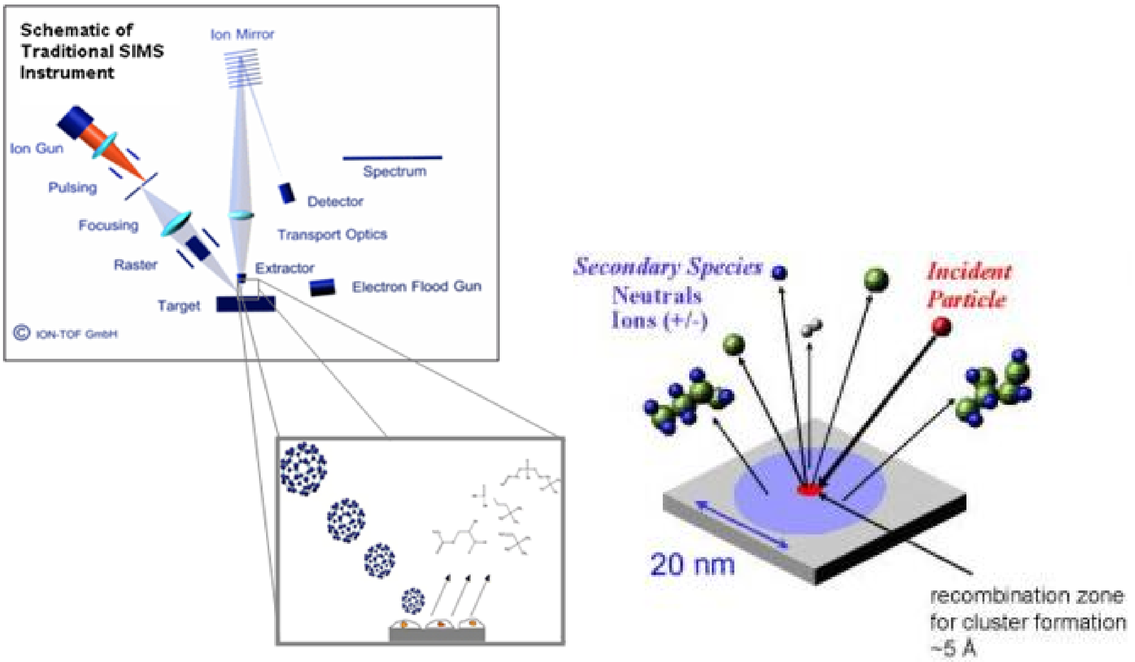
1.3 Fundamental Principles

The SIMS equation as given by [79]:

$$I_m = I_p Y_m \alpha^+ \theta_m \eta \tag{1.1}$$

Where I_m is the secondary ion current of species m, I_p is the primary particle flux, Y_m is the sputter yield, α^+ is the ionisation probability for positive ions, θ_m is the fractional concentration of m on the surface layer and η is the transmission of the analysis system. The SIMS equation gives a framework for understanding the process of the experimentation.

Samples are prepared in an ultra high vacuum. Primary ions are accelerated through an electrical field toward the sample stage. Kinetic energy, and thus particle velocity, are generated through the relationship of $Ek = qV = \frac{1}{2}mv^2$. The beam of primary ions impact the surface with energies in the range of $10s$ of keV. The primary ion beam causes a collision cascade amongst surface atoms and atoms, molecules and molecular fragments are usually ejected from the material's surface. Secondary particles produced closer to the site of impact tend to be dissociated ions (positive or negative). Secondary particles generated farther from the impact site tend to be molecular compounds, typically fragments of much larger macromolecules [65]. The secondary ion yield depends on the nature of the analyte and has become an important value to understand, especially when trying to image a sample. These values must be known to reconstruct an accurate quantitative description of the



(a) Diagram representing the SIMS process. Reproduced from [26]

(b) Diagram representing the impact site.

Figure 1.1: Schema of the SIMS process.

original sample chemistry.

The secondary ions are then accelerated into a flight path on their way towards a detector. Because it is possible to measure the “time-of-flight” of the ions from the time of impact to detector on a scale of nanoseconds, it is possible to produce a mass resolution in sub atomic mass units. Figure 1.1 gives an diagrammatic overview of this process. By analysing the observed spectra in terms of fragmentation and subsequent ion formation of the sample this can reveal the overall structure or composition of the original sample.

1.3.1 Static SIMS

As mentioned above Static SIMS is a regime which imparts a low primary ion current to hit ‘fresh’ regions of the sample surface to liberate ions, molecules and molecular clusters for analysis. The conditions require a very low primary ion dose (less than $10^{13}cm^{-2}$ ions for a gallium beam), which is much less than 1% of the surface layer that is impacted by the primary ions [79]. This means a small amount of primary ions are used to bombard the sample per area per unit time. Static SIMS is used to determine surface concentrations of elements and molecules without significantly altering the sample. Static conditions are defined as those which maintain the integrity of the surface layer within the time-scale of the experiment.

1.3.2 Dynamic SIMS

In contrast to static SIMS, “dynamic” SIMS is the method of choice for quantitative analysis, because a higher primary ion current results in a faster sputtering rate and produces a much higher ion yield. Thus, dynamic SIMS gives a higher secondary ion response. TOF-SIMS is also capable of sputter depth profiling. The higher ion flux used in dynamic SIMS eats away at the surface of the analyte, the beam steadily ‘digs’ deeper into the sample and generating secondary ions that characterise the composition of the sample as a function of depth. Depth profiling allows monitoring of all species of interest simultaneously, and with high mass resolution. By monitoring the entire mass spectrum at each depth unexpected contamination in the sample can be detected.

1.3.3 Primary Ions

For SIMS analysis there are a large number of properties to consider when choosing a primary ion source. The primary ion source of choice depends on the goal of the analysis. First generation primary ions were inert gas ions (Ne^+, Ar^+ and Xe^+) [9]. These sources were then followed by surface ionization sources, typically Cs^+ ions. liquid metal ion source (LMIS) generated ions used were, and are, Ga^+ and In^+ [3].

It was proposed by [5] that if clusters break up on impact then many individual atoms will bombard the surface of an analyte almost simultaneously and this would promote larger fragments from the surface. This has been shown to be the case in numerous studies with different cluster primary ion beams such as [46]. It was also shown by molecular dynamics simulations comparing Ga^+ and C_{60}^+ primary ions [56] [58] that cluster ions provide more secondary ions and more high mass ions. The use of buckminsterfullerene (C_{60}^{+}) is a quite recent addition to the spectrum of ion sources, offering high secondary ion formation efficiency and low sub-surface chemical damage. The low chemical sub-surface damage by C_{60}^{+} primary ions makes molecular depth profiling and analysis far beyond the static limit possible [24] [81]. Low chemical damage and high secondary ion yields have also been shown for even larger cluster primary ions like Au_{400}^{4+}[45] and more recently $Ar_{1000-5000}^{+}$ [1] and $(H_2O)_{1000}^{+}$ [67].

1.3.4 Detectors

There are generally two detector devices that are used to process the signal in SIMS instrumentation, a time to digital converter (TDC) and an Analogue to Digital Converter (ADC). Classical SIMS instruments generally use a TDC, this leaves them subject to Poisson noise due to the counting statistics. As it is a digital system, multiple simultaneous events cannot be recognised, thus a finite dead time can be observed. Therefore, with high secondary ion fluxes all ion signal may not be recorded. With recent SIMS experiment this has become an apparent issue. Though Poisson noise can be scaled out with a high degree of accuracy [42] and recently dead time corrections have been implemented [76].

An ADC is used in the J105, also the Biotof can select the use of an ADC e.g. for post-ionisation experiments. The ADC directly maps the electron cascade as opposed to the TDC which counts binary when the voltage exceeds a threshold. This leaves the J105 outside of counting statistics and thus Poisson scaling.

In an ideal world SIMS detection would result in pin-sharp peaks. However, there are issues of ions with different sputter energies and their associated flight paths through the analyser that give rise to a 'peak width' and 'peak shape'. The analyser type will have an effect on both the shape and width.

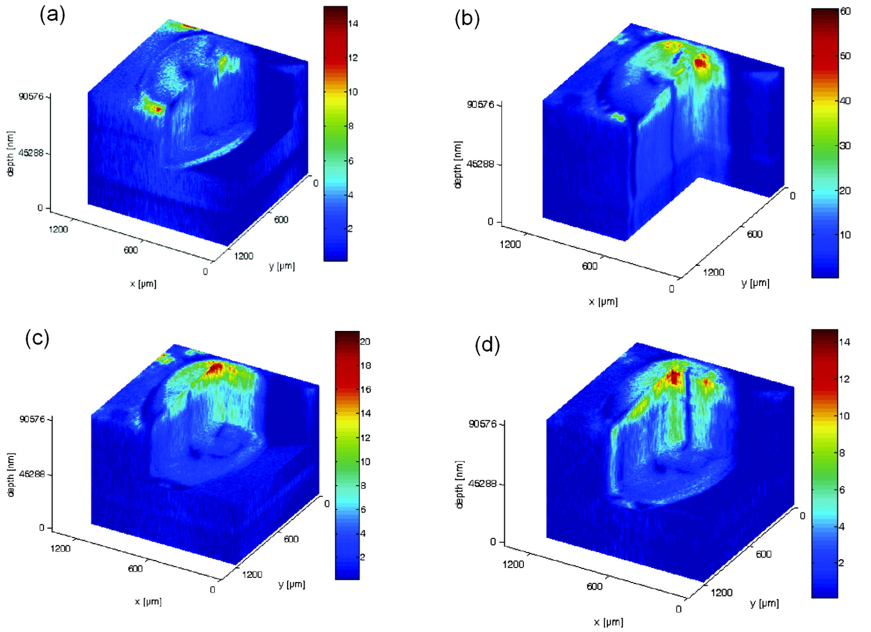


Figure 1.2: 3D biochemical images of freeze-dried oocyte, showing changes in (a) phosphocholine peaks m/z 58, 86, 166, and 184, (b) signal summed over the m/z range 540-650, (c) signal summed over the m/z range 815-960, and (d) cholesterol peak at m/z 369. Color scale normalized for total counts per pixel for each variable (m/z range) [23].

1.3.5 Image data

In imaging mass spectrometry, images are created based on mass spectral information. Ions are generated from the surface of the material as the primary ion beam is rastered across the sample surface. The generated ions are then detected in a mass analyzer. This is done for an image of e.g. 256x256 pixels. Spatial resolution is dependent on the primary ion beam widths, ranging from about 50 nm to 1 μm. The resulting dataset contains a full mass spectrum for each image pixel. These mass spectra can contain intensity values for individual, multiple or a whole range of m/z values.

The total ion image is the sum of all peaks in the spectrum, this can be viewed to give a topographic map. When a peak of interest is observed by the analyst, an intensity plot can be made that visualises the distribution of the related ion at the sample surface and used to interrogate regions of interest, for their chemical composition, via computer processing after the datasets have been instrumentally acquired. When this is done for multiple m/z values of interest, each m/z value can be assigned a colour, giving a colour plot that visualizes the spatial distribution of several ions throughout the sample. The acquisition of a full mass spectrum at multiple image pixels results in a “data-cube”, a collection of intensities for each x-position, y-position and m/z value. The choice of data storage during or after acquisition depends on the way data were acquired. By imaging multiple layers of a sample a 3D model can be constructed from the data which allows for spatial distribution to be analysed in three dimensions [21].

The ability to analyse samples in such a manner has drawn great attention in recent times due to its applicability to biological samples [46]. Figure 1.2 shows a 3D representation of a *Xenopus laevis Oocytes* as imaged by ToF-SIMS [23]. Visualisation of SIMS data is discussed more in Section 5.3.

1.3.6 Approaches to Data

Raw data from the mass spectrometers comes in the form of mass spectra with attached meta-data. Depending on the resolution at which the data are acquired, the size of the files will vary up to very large files greater than tens of gigabytes. There are currently two proprietary formats in-house, one for each instrument.

One current impediment in the ToF-SIMS community is that of file formats. Currently there are a few manufacturers of ToF-SIMS instrument (ION-TOF[41] and PHI [57] to name two). With each of these comes a proprietary file format. Generally these file formats are not open, therefore data transfer becomes a problem. Beside data transfer, problems due to file formats no longer being supported can also arise. Though there is a standard, it is rarely used. Proprietary file formats can be a problem for analysing or transferring data. Standardisation of file formats is an important step to integrating a community of research. This along with the lack of large databases hinders the potential of ToF-SIMS.

Interpretation of individual spectra can be characterised in two ways. The first is interpretation by comparison with other spectra. The second is by using fragmentation patterns from molecular information. Thus, identifying molecular structure prediction from the spectral data. Complementary data from other techniques can also be an aid in the process. This thesis is designed to simplify some of the tasks associated with these interpretive steps.

In mass spectrometry mass spectra contain all the molecular/atomic information. By examining these, the researchers can determine the composition of the test sample. However, datasets from the instrumentation have become enormous. To deal with these massive datasets statistical methods have been applied to aid users in their analysis of these datasets. Multivariate Analysis (MVA) tools can be used to extract information from these. Methods such as Principal Component Analysis (PCA), Discriminant function Analysis (DFA) [70], Partial Least Squares (PLS) [33], Multivariate Curve Resolution (MCR) [75], Maximum Autocorrelation Factor (MAF) [77] and others have been utilized. Especially when dealing with large data sets, tools like PCA can be invaluable.

The goal of using these methods is to extract information from the datasets which may be unattainable by conventional methods. PCA is most commonly used to extract differences between samples or datasets and thus classify them into groups. PCA has also been used to extract chemical information for imaging. DFA has often been applied as a subsequent technique to PCA for further discrimination [2].

PCA is an unsupervised method for identifying patterns in data. The goal of PCA analysis is to identify trends or patterns, through variance, in data and expressing the data in terms of Principal Components (PCs), which are axes with the maximum variance. The first principal component (PC1) is the direction that describes the highest degree of variance through the data. The second principal component (PC2) is the axis in a direction describing the highest degree of remaining variance after the first PC and is orthogonal to it.

Classification algorithms have begun to become more prominent in their applications to SIMS data. classification routines will be discussed in more depth in Section 4.

1.4 Aims

The primary aim of this thesis is to improve current methods for dealing with SIMS data. To do this, two key areas relating to SIMS data are identified:

- The very large size of current SIMS images/image stacks.
- The complexity of the data.

These two overarching areas can be broken down into sub-categories, some of which this thesis focuses on. When discussing file size issues such as storage, computer memory and processing the data crop up. If trying to comprehend the complexity of data many approaches can be taken, such as simplification of the data, incorporating complex algorithms for MVA and approaches for intuitive visualisation. To broach these issues some novel approaches to SIMS data are proposed.

From here the thesis takes the following form:

In Chapter 2 a new algorithm for automatic peak picking of SIMS data is introduced and tested.

In Chapter 3 some fresh techniques for compression of SIMS images are presented and evaluated.

Chapter 4 introduces new classification routines to SIMS data and compares the results to a commonly used approach of PC-DFA.

There are two major sections in Chapter 5, the first section brings new procedures for reading files and mean centring of sparse data for ease of analysis. Some spectral matching tools are also introduced here as an aid in spectral interpretation. Finally a collection of tools created in Matlab are introduced to automate standard processing techniques such as visualisation and performing PCA on data.

Finally in Chapter 6 there is some general discussion regarding the outcomes of the experiments and approaches presented and some final conclusions are drawn.

Chapter 2

Peak Picking Applied to ToF-SIMS

Contents

2.1 Introduction

The term 'peak picking' is used to describe the exercise of determining a 'discrete spectrum from continuous data'[†]. This is in contrast to arbitrary peak selection based on *a priori* knowledge of the sample [80] and/or by applying a simple intensity threshold to the data.

Peak picking fundamentally is a method of locating peaks within a spectrum. Classically peak picking has been used to find peaks to generate peak lists [74]. From this the area of each peak discovered can be used instead of its original or fitted distribution, thereby reducing the information to a single mass channel. There have been many approaches adopted to do this by the mass spectrometry community [44][54].

Peak detection is a challenging task since mass spectra are often tainted by noise, ToF-SIMS is no exception to this. As a result, various algorithms have been proposed to facilitate the identification of informative peaks. These algorithms differ from each other in their principles, implementations and performance [83]. In order to create a functional peak picking algorithm for ToF-SIMS some open source algorithms are investigated in relation to SIMS data.

ToF-SIMS has gone through a resurgence in recent years with the ability to do 3D biological chemical analysis. With this, data which are much more complex and vastly more complicated to interpret with respect to a sample has also followed. To understand these data sophisticated computational approaches are needed. Reducing the complexity of data by peak picking can be viewed as one of these approaches. For peak picking to be an option as a pre-processing technique, reliable, reproducible and accurate methods are imperative. In this chapter the aim is to provide a comprehensive comparison of existing peak detection algorithms and extract reasonable principles and schemes on which to apply them.

It is possible to peak pick manually but aside from the time aspect, the quantity

[†] A term suggested at the 59th IUVSTA Workshop on *Surface Chemical Analysis: Improving data interpretation by multivariate and informatics techniques*, Trinidad and Tobago, 2010

and the necessary accuracy make this unreasonable, therefore it is recommended to do it automatically. For a human, detection of most peaks is not an issue. Intuitively our eyes are able to discover the largest peaks with a quick glance, but generally with less accuracy, due to some subjectivity, when concerning smaller peaks. Automatic peak detection is a more robust approach, as the same algorithm will produce the same results, a fact which cannot be said of manual peak selection. However, detecting true signal from a spectrum still remains an issue due to low abundance signal, which may be buried by noise. This can cause high false positive rates of peak detection. Overlapping peaks or shoulder peaks can corrupt interpretation of a spectrum, distorted peaks due to acquisition characteristics or asymmetries in peaks can cause errors in detection rates which all affect the efficacy of peak picking approaches.

2.1.1 What is a peak?

To expand on the process of peak picking a definition of what a peak actually is is needed. One commonly used definition, initially developed by Yasui *et al.* [84], relies on local maxima, that is to say a point with higher intensity than N other points in a local neighbourhood. This approach applied to peak picking is generally performed by searching for maxima in a pre-defined window width, repeating the process so that window covers the entire spectrum. A list of potential peaks is returned. Window widths which are too small will result in keeping low intensity noise fluctuations as peaks, while values that are too big will potentially exclude true signal. Another common way of considering a peak is in relation to signal-to-noise ratio. Here, a peak is defined as signal that is observed above three times a common baseline for example.

These definitions are simplistic and do not deal with more complex spectral features such as overlapping peaks or shoulder peaks. More information is available in a spectrum other than just a maximum, in regards to spectral features/peaks. Every peak has an underlying distribution dependent on the type of machine and the settings of that machine. Therefore, a peak can be defined as a clear signal distribution above an observed noise level or baseline where an underlying distribution can be observed. This definition is general enough to cover all types of mass spectrometry but specific enough to discriminate peaks from noise. This definition does not however deal with the stated fact of overlapping or shoulder peaks. Following this, an overlapping peak is defined as a distribution where two or more evident peaks are observed without the ability to resolve either peaks underlying distribution clearly or fully. A shoulder peak is defined as a peak overlapping with another peak to the extent that a clear valley between the two peaks is no longer evident, but there is an obvious deviation from the peaks expected underlying distribution due to an additional signal distribution.

2.1.2 What can Peak Picking be used for in ToF-SIMS?

Classically peak picking has been used for single spectrum analysis or collections of spectra. From this peak lists can be generated. These peak lists can then be used in a variety of ways such as databasing [69]. By peak picking, fitting can be performed and more information about a peak can be gleaned. The area and standard deviation of each peak can be calculated and used instead of its spectral intensity. Thus, peaks can be reduced to a single mass channels. In other fields of MS, peak picking has been an input in Biomarker discovery [85]. In Chapter 3 peak picking is used as a tool for compression of SIMS images, as opposed to manual peak selection.

2.2 Peak Picking routines

Recently in the MS community it has become apparent that more statistically relevant methods for automatically finding peaks are a necessity. In the mass spectrometry community as a whole there have been reviews of some of these algorithms [83][85]. By incorporating aspects of several approaches a new peak picking algorithm tailored for ToF-SIMS is proposed here. Due to the nature of the SIMS process and its resultant spectra, in comparison with other MS techniques, a tailored peak picking algorithm is needed which can accommodate its nuances. Features such as the large disparity in signal-to-noise ratios, and signal intensity as a whole, across a spectrum and the high levels of peak overlaps need to be accommodated to produce an accurate representation of a spectrum.

This method however can easily be adopted for other forms of MS or indeed any data with Gaussianesque peaks, where the goal is to locate peaks. Generally peak picking algorithms take the form of baseline estimation, noise smoothing or removal and peak picking, a brief overview of peak picking is given below.

Baseline

Baseline correction is typically a two-step process, estimating the baseline and subtracting the baseline from the signal. However, unlike other forms of MS where baseline correction is needed, in ToF-SIMS it is not necessary. This is due partially to the low levels of chemical noise in SIMS, in comparison to techniques such as MALDI. However, noise is still a problem when it comes to automated peak identification in SIMS.

Smoothing filters/ noise removal

Smoothing filters are often applied as a pre-processing step in peak picking routines, one type of these used is a moving average filter. A moving mean average filter is used to smooth out high frequency signal while retaining lower frequency signal. This is achieved by taking the average of N neighbouring points (user defined value) of data and substituting the averaged value for the value at that point. Once this is done the next point in the series or spectrum is calculated in the same way.

Saviztky-Golay [63] filtering can be considered as a generalized moving average filter. It performs a least squares fit of a small set of consecutive data points to a polynomial and takes the central point of the fitted polynomial curve as output.

This type of filter has been widely adopted. Wavelet filtering has also become a popular approach for noise removal [53][51].

When smoothing is applied to mass spectra it smooths out noise but it can also smooth out real spectral features such as low intensity peaks and shoulders. If real signal is smoothed out and consequently not detected by the algorithm this information is lost in future analysis that uses this approach.

Peak picking

One of the most basic approaches adopted was that of Yasui *et al.* [84]. A peak was found by scanning through a spectrum and evaluating each point. If the point is a local maximum of N neighbouring data points, where N is some arbitrary number of data points, then this is considered a peak. This method, though being simplistic, is effective if the peaks are free from noise, noise has been removed by smoothing and/or noise removal techniques have been adopted. Since noise is ever present in SIMS data this brings into question the efficacy of the noise removal and smoothing filters that are implemented. More recent approaches have negated the used of noise filter by using wavelets, this is discussed more below, Section 2.2.2.

Intensity threshold

A simple solution to the problem of peak picking with noise is to use an intensity threshold. By setting some arbitrary value as an intensity threshold all points above the threshold can be considered as peaks. Then applying smoothing and a local maximum function as above gives an estimate of the peaks in the spectrum. However this does not take into account true peaks below the intensity threshold.

2.2.1 Cromwell

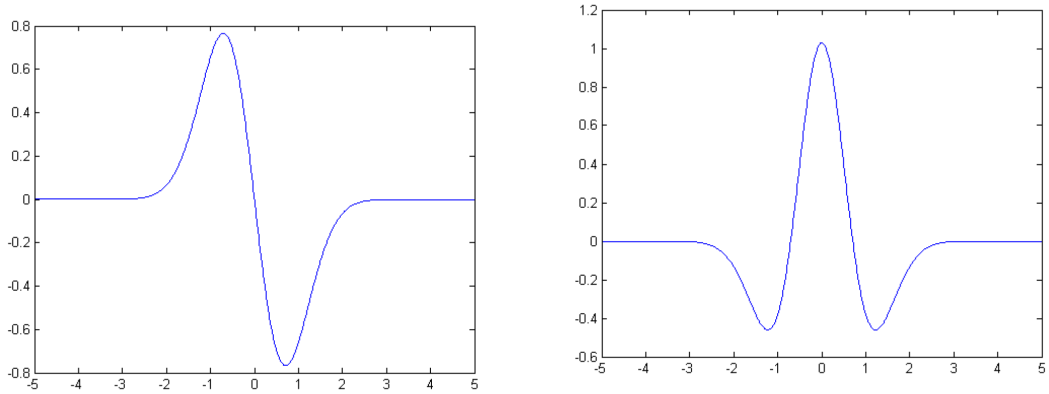
Coombes *et al.* [13] created a peak picking approach which was an extension to simple locating local maxima in a spectrum. They applied their approach to MALDI data. Their method relied mainly on pre-processing of spectra using calibration, de-noising, baseline removal, noise estimation and then locating local maxima. By performing these extensive transformations to their data they aim to remove the difficulties associated with searching for local maxima. They investigate two approaches, one using single spectra the other using the mean spectrum.

Firstly they ensure that all spectra are well calibrated. Then they de-noise the spectra using wavelet regression, the Undecimated Discrete Wavelet Transform (UDWT). Hard thresholding was performed in their case to remove noise. The wavelet coefficients are calculated and all coefficients less than a certain threshold are set to zero. An estimate of the noise level across the spectrum was calculated by using a median filter. Estimation and removal of the baseline is then performed on the denoised signal via a monotone local minimum curve calculation. The spectrum is then normalised to total ion counts. Peaks are then found via a local maximum search and a signal to noise ratio calculated. Peaks below a certain signal to noise ratio are discarded.

This method applies steps which may not be applicable to SIMS data such as baseline removal, as there is no baseline in SIMS data. Also the use of a denoising step is negated in the Continuous Wavelet Transform (CWT) approach of Du *et al.* [17]. which will be discussed further in the coming sections. This method will be evaluated against other peak picking routines on SIMS data in the Results, Section 2.5.

2.2.2 CWT

Du *et al.* [17] introduced the concept of using the continuous wavelet transform for the purpose of peak picking. This method has become very successful in the mass spectrometry community [83]. Peak picking is carried out by applying the continuous wavelet transform to a spectrum. This is done through implementing the Mexican hat wavelet (Figure 2.1b) which is proportional to the second derivative of a Gaussian probability density function. This produces a 2D coefficient matrix with size of M x N (Figure 2.3d), where M is the number of scales of the wavelet



(a) First derviative Gaussian wavelet.

(b) Second derviative Gaussian (Mexican hat) wavelet.

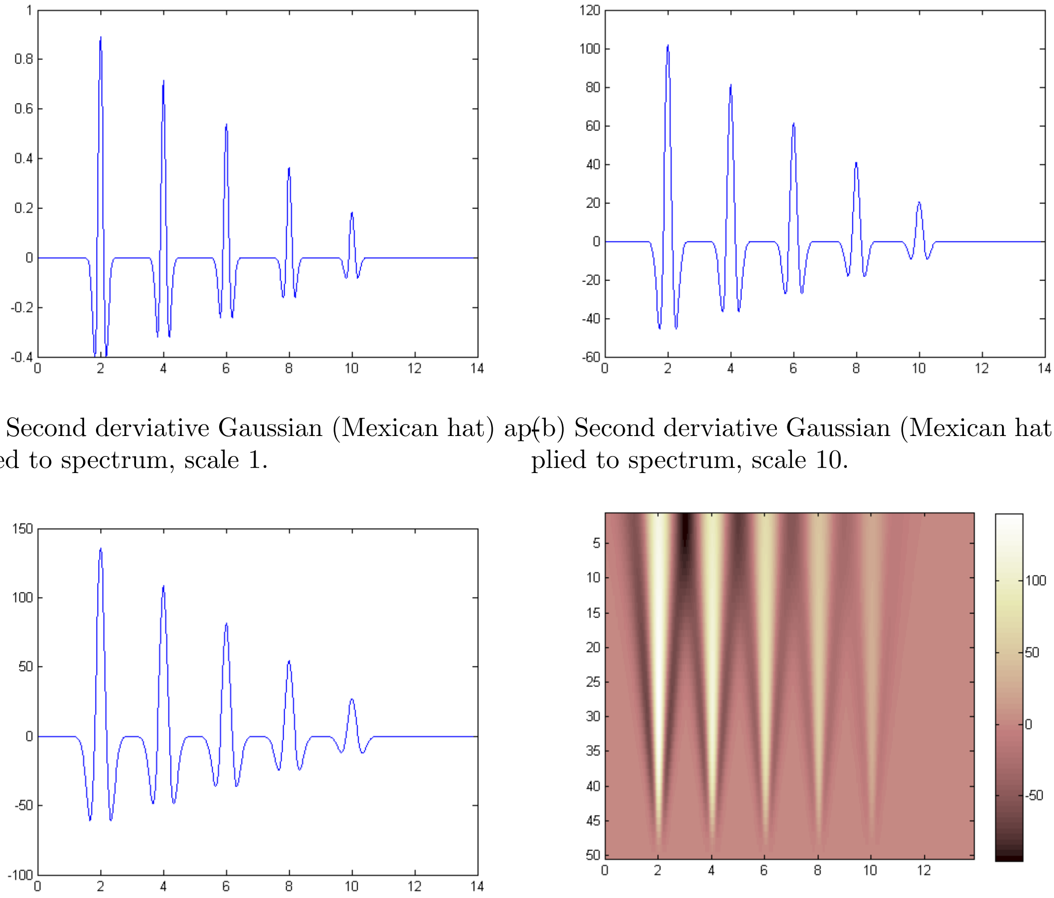
Figure 2.1: Graphical representation of the first and second Gaussian derivative wavelets.

and N is the number of mass channels in the spectrum.

To illucidate the process the Mexican hat wavelet was applied to a simple model spectrum from Figure 2.5a. Figure 2.2 shows the results of this. Figure 2.2a-2.2c show the wavelet coefficients for the scale 1, 10 and 15 respectively. Figure 2.2d is the 2D plot across 50 wavelet scales.

The Mexican hat wavelet with scale one provides the best match (highest amplitude in the cwt coefficients) for peaks with a width of about two mass channels. The amplitude of the cwt coefficients gradually increase as the width of a peak increases, reaching a maximum when the scale best matches the peak width, and gradually decreases as the scales continues to increase. Viewing this as a 2D plot (Figure 2.3d) ridges' corresponding to the peak locations can be observed as the scaling and shifting of the wavelet change. These can be viewed as ridges in 3D space (Figure 2.16b is an example of this) and reduces the problem of peak identification to finding the maxima in the cwt coefficients and linking them as ridgelines, which correspond to peak locations, in this 3D space. One of the advantages of this approach is that by using the cwt transform a denoising or smoothing step is not required as wavelets are invariant to noise.

The peak picking process outlined by Du *et al.* is to first find the local maxima



(a) Second derviative Gaussian (Mexican hat) applied to spectrum, scale 1.

(b) Second derviative Gaussian (Mexican hat) applied to spectrum, scale 10.

(c) Second derviative Gaussian (Mexican hat) applied to spectrum, scale 15.

(d) 2D coefficients image of second derivative Gaussian (Mexican hat) wavelet applied to spectrum.

Figure 2.2: Second derviative Gaussian (Mexican hat) wavelet applied to a simple spectrum.

at each wavelet scale. A sliding window is used whose size is proportional to the wavelet scale to find these maxima. The next step is to link these local maxima as lines, “ridge lines”. To create a ridge line, each maximum is found at the largest wavelet scale, then the nearest maximum point in the next scale directly below is investigated. The distance between two adjacent points on a ridge line should be smaller than some window/threshold. If there is a maximum within the window scale, this point is added to the ridge line and the next scale down can be searched for a maximum. If the gap is larger than a given threshold, the ridge line is terminated. At the end of each ridge line the points on the ridge are removed from the searching list so as not to interfere with future ridges. The search is then re-initiated on the rest of the maxima in that scale until all maxima at that scale have been searched and removed. Following this, searching of the remaining maxima in the next lower down scale is carried out until all maxima are exhausted from the cwt coefficient matrix at the smallest scale.

Having created all the ridge lines, ridge lines with a length larger than a certain threshold are saved with ridge lines shorter than this threshold discarded and not used for further analysis. Ridge lines are beneficial as false peaks can be removed if the length of their ridge line is smaller than the given threshold set by the users. Du *et al.* estimate the width of a peak as being proportional to the scale corresponding to the maximum amplitude on the ridge line, this allows for a peak candidate to be dropped if its width is not in a given range.

To identify the peaks based on the ridge lines three rules were defined to identify the major peaks:

1. The scale corresponding to the maximum amplitude on a ridge line is proportional to the width of the peak; therefore this should be within a certain range corresponding to a tolerance allowed for the width of a peak.

2. The signal-to-noise ratio should be larger than a threshold.

3. The length of ridge lines should be larger than some threshold.

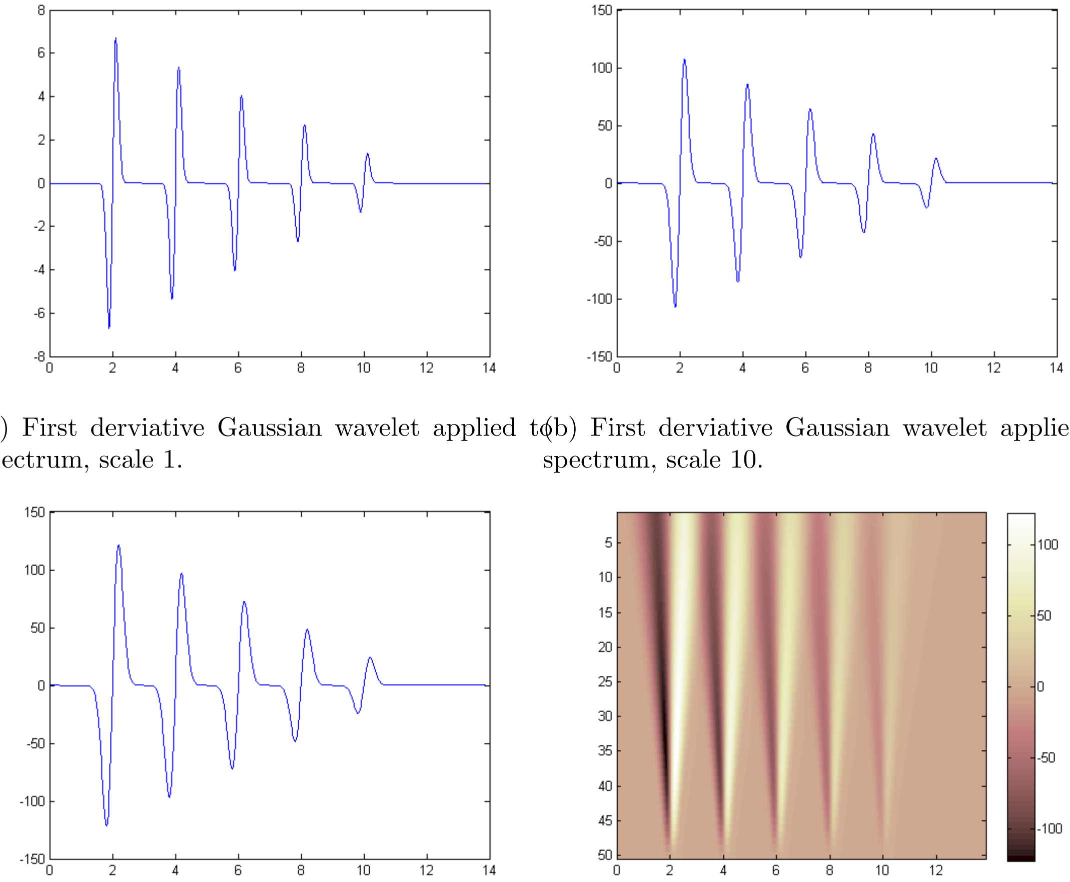
For a broader introduction to wavelets and their advantages over Fourier transforms for spectroscopic data see [12].

2.2.3 GDwavelet

Nguyen *et al.* produced a peak picking algorithm also based on wavelets called GDWavelet [53]. GDWavelet utilizes three main processes: a bivariate smoothing model, Gaussian derivative wavelets, and envelope analysis. Gaussian derivative wavelets are investigate from the first Gaussian derivative, the Mexican hat (second derivative) and the third derivative. Nguyen *et al.* approach the problem slightly differently than Du *et al.* investigating three derivatives, ultimately only using the first two for analysis. They also make extensions to the approach using bivariate shrinkage for denoising and envelope analysis to retain small amplitude ‘real’ peaks.

The use of the bivariate shrinkage function is to remove noise without removing true signal. This is carried out in the Stationary wavelet transform (SWT) domain. Research by Sendur *et al.* [66] showed that algorithms utilizing the dependency between coefficients can give better results than those using the independency assumption. Therefore they created a smoothing function which employs the dependency across coefficients/scales.

In their paper Nguyen *et al.* propose using the zeroline crossing lines instead of ridgelines of maxima. Zeroline crossing is as it sounds, the point at which signal crosses zero. They achieve this using the first derivative Gaussian wavelet, Figure 2.1a. As with Du *et al.* this is performed in the wavelet domain. For the first derivative of a Gaussian wavelet applied to MS type data this generates one zeroline crossing for a peak. Figure 2.3 shows the results of applying the first derivative Gaussian wavelet to the same simple spectrum as Figure 2.2. The Mexican hat generates two zeroline crossings and for the third derivative this generates three zeroline crossings. Nguyen *et al.* use the first derivative zeroline crossing to find peak centres, following this, the second derivative (Mexican hat) is used to estimate the peaks parameters.



(a) First derviative Gaussian wavelet applied to spectrum, scale 1.

(b) First derviative Gaussian wavelet applied to spectrum, scale 10.

(c) First derviative Gaussian wavelet applied to spectrum, scale 15.

(d) 2D coefficients image of first derivative Gaussian wavelet applied to spectrum.

Figure 2.3: First derivative Gaussian wavelet applied to a simple spectrum.

To find a peak's centre they apply the wavelet transform using the first derivative of a Gaussian. They then find all zerocrossing points. By linking these points across all scales, similar to linking ridgelines, in the wavelet domain the peaks centre can be estimated. Since there are no maximum on a zerocrossing line they use the average position of the zerocrossing line as the centre of the peak. Given by:

$$\mu_i = \frac{1}{N}\sum_{s=1}^{N} u_o(s) \tag{2.1}$$

Where μ_i denotes the peaks centre, $u_0(s)$ is the zerocrossing line at scale s, for a line of length N. To estimate the standard deviation of a peak, they use:

$$\sigma_{i-left}(s) = \sqrt{(u_{0left}(s) - \mu_i)^2 - \frac{s^2}{2}} \tag{2.2}$$

$$\sigma_{i-right}(s) = \sqrt{(u_{0right}(s) - \mu_i)^2 - \frac{s^2}{2}} \tag{2.3}$$

Where $\sigma_{i-left}(s)$ is the estimated standard deviation for the left side of the peak at scale s, $u_{0left}(s)$ is the zerocrossing to the left at scale s, u_i is the estimated centre and s is the scale. These values can then be calculated across all scales, by averaging these values an estimate of the standard deviation can be derived.

In their paper they also go on to propose a method for estimating the peaks true height using the above calculated values.

Zerocrossing lines are a similar approach to ridgeline analysis in that zerocrossing lines below a certain threshold can be discarded. For an in-depth look at wavelets see [29]. This method is evaluated against the CWT method outlined above in the Results, Section 2.5.

2.3 Peak Picking Routine for SIMS data

In order to perform accurate peak picking on SIMS data a comparison was carried out on the core processes of two of the discussed methods above, the CWT approach of Du *et al.* [17] and GDWavelet as adopted by Nguyen *et al.* [53]. The results of this comparison can be seen in Section 2.5.1. The results of this section inform the rationale for adopting the approach outline below.

The philosophy adopted when designing the peak picking approach was to be as accurate as possible in finding peaks, overlapping peaks and peak shoulders. The inclusion of noise peaks can be a problem with such algorithms, hence a well designed algorithm must exclude as much erroneous data while incorporating maximal true information. This includes peak shoulders. Finding of and separating peak shoulders can add real chemical information about a particular sample. To achieve this, the peak picking algorithm must not only find peaks and shoulders but also separate/deconvolve them. This is necessary since overlapping peaks can distort the desired information in neighbouring peaks and peak shoulders are also an example of this. An estimate of the actual information contained in these shoulders and overlapping peaks is essential to obtain an accurate representation of the chemistry contained within a mass spectrum. Below is an outline of the approach that was found to retain the most accurate information about a spectrum while discarding unwanted noise.

2.3.1 Finding spectral features

Following the approach of Du *et al.* peaks were found using the continuous wavelet transform via the Mexican hat wavelet. When the cwt is applied to a spectrum the problem of finding a peak is transformed into finding maxima in the cwt coefficients. The use of ridgelines allows the exclusion of small unwanted peaks and noise. Furthermore, the setting of a minimum threshold ensures that low resonating data are excluded from ridgelines. The cwt approach is compared to the zeroline crossing method in the Results Section 2.5.1.

As has been discussed previously, SIMS data contains a wealth of information. Some of this information comes in the form of overlapping/shoulder peaks which can contain instructive chemical information. It is desirable to know if these peaks are present. To find shoulder peaks the peak picking routine needs to be kept highly sensitive in terms of scale and ridge length. This may force the algorithm to incorporate erroneous peaks. However, these can generally be filtered out using the noise peak removal steps below.

Following the analysis in the results section it becomes apparent that the ridgeline approach can fall subject to picking the wrong peak centre due to overlap or shoulder peaks. To combat this, the zeroline crossing method is applied as an additional check of the peak's centre. The zeroline crossing can fall prey to similar effects as the cwt. However, if operated over small, therefore sensitive, scales the effect due to overlap or shoulder peaks should be small if non-existent.

Thus, the cwt approach is applied to the spectrum to find peaks and shoulders, then the zeroline crossing is applied to adjust peak centres. If there are no zeroline peak centres within a local region then the cwt centre is used, if there are multiple then the closest is used.

2.3.2 Mass scale variance

Peaks widths can vary greatly in terms of standard deviation and in terms of the number of data points they encompass across the mass spectrum. Generally peaks get broader at higher mass values due to the non-linear ToF detection. To facilitate this, different wavelet scales can be searched for different mass regions. By increasing the starting-finishing wavelet scale range the ridgeline approach becomes less sensitive to small peaks and noise. It is important to include the scale which contains the maximum on the ridgeline. Also, this is a filtering method to exclude noise.

As the width of peaks being searched increase in terms of their position on the mass scale the wavelet scale range is also increased in keeping with the increases in peak width along the mass scale. This is discussed further in the Results, Section 2.4.1.

2.3.3 Parameter estimation

Parameter estimation as adopted previously by Du *et al.* will suffer from the same problems as described above with overlap in larger wavelet scales. This can include the maximum coefficient on the ridgeline. This subsequently affects outcome of their parameter estimation. The methodology of Nguyen *et al.* has a more robust interpretation. They recommended to take an average across all scales using the zeroline crossing of the Mexican hat. They use the zeroline crossing at both sides of a peak at each wavelet scale. Thus this approach is more robust to noise and fluctuations. It is noted that there is still the problem of overlap when two peaks are close or the scales chosen are larger than necessary. Consequently, it is proposed that only the first several scales are used to compute a value for standard deviation or FWHM. This reduces or eliminates the problem of the overlapping signals causing an effect in the wavelet domain and gives a much better estimation of the peaks true values.

By averaging across wavelet scales this can reduce error due to overlapping peaks, large noise or low signal-to-noise ratio. However, accurate information is need for peak picking to be implemented in the application proposed in the Chapter 3, as one of the goals is to separate overlapping signal in an image, therefore peaks are fitted to the data (outlined in Section 2.3.5).

2.3.4 Noise/erroneous peak removal

Classically to remove unwanted noise smoothing techniques have been employed. The goal to only remove noise and noise peaks and not to remove real signal. Given the information gathered from the previous steps numerous thresholds can be set to remove unwanted noise peaks without smoothing of data. These parameters can be tuned as desired by the user depending on the level of accuracy or compression desired.

As was suggested by Du *et al.* a signal-to-noise ratio can be used. Here all peaks that have been picked become the numerator with noise in the local region as the denominator. Noise is estimated in the wavelet domain as the average value of the coefficients in the first wavelet scale in the local region. Peaks with a ratio below a threshold can then be excluded. Du *et al.* suggest a SNR of three. In practice this value works well. However, for data such as J105 data where noise is more monotonic then in conventional machines, this value can be too high and exclude subtle peaks as a result, thus this parameter should be lowered or excluded.

Using the estimated peak values from the previous section of parameter estimation or by using the values from the peak fitting section (Section 2.3.5), peaks can be excluded due to their estimated FWHM or their standard deviation. Here a minimum/maximum threshold is again applied as the tunable parameter. If the peak does not meet this threshold then peak can be excluded. This threshold should be carefully set as peaks can vary widely across a spectrum. Different values for different mass ranges can also be set.

Finally a simple intensity threshold can also be set. For the analyst who only wishes to observe peaks of certain intensities, a minimum threshold can be set to exclude peaks that are too low in intensity. In using this parameter caution should be observed as noise levels generally vary across the mass spectrum and a simple threshold may be too simple an approach as low intensity peaks at high mass may be characteristic of the sample. Again, a threshold can be varied according to mass range.

2.3.5 Peak fitting

Once the peaks parameters have been estimated peak fitting routines can be applied. Here, scripts written in MATLAB (version R2009a, MathWorks Inc., MA, USA), some utilising routines from the PLS_Toolbox (version 4, Eigenvector Research Inc., WA, USA), were used to fit Gaussian peak shapes to spectral features discovered in the spectrum. The Gaussian peak shape was adopted in order to exemplify the procedure performed. Other mathematical functions, or peak shapes derived directly from the data, are equally applicable. Once fitted, the resulting components provide the underlying peak positions, widths etc. These computed values can be used for further calculations or as a guide to the data.

A least squared fitting routine was adopted to fit the peaks. Here upper and lower bounds need to be set to constrain the algorithm. These bounds were set loosely because, in general, peaks values and location can be estimated to a reasonable degree. However, there can be some outliers in this approach, discussed more in Section 3.5. Also it should be noted that to optimise the whole spectrum at once takes far too long. Thus, sections of the spectrum are assessed in separate regions. This was done to decrease the time to compute the result. This can cause peaks to be fitted incorrectly, discussed more in Section 3.5.

2.4 Data

2.4.1 Model Data

Model data were generated to exemplify the strengths and weaknesses that an approach may have. By generating data with a wide variety of features that are likely to occur in real data a judgement can be made about an approaches efficacy.

Model noise

Noise is created by a call to the *randn* function in Matlab which creates normally distributed pseudorandom numbers. These then can also be controlled and varied by adding a mean and multiplying by a standard deviation. The noise can be more or less intense by varying these values. This noise can then be added to the subsequent model data for further analysis of the algorithms accuracy. An example of noise with a mean of 1 and a standard deviation of 0.5 can be seen in Figure 2.4. eight sets of noise were created by increasing the standard deviation in 0.5 increments from 0.5 to 4. These varying noise levels were then added to the model spectra to create the effect of increasing the signal-to-noise ratio similar to the drop in signal-to-noise in real SIMS data.

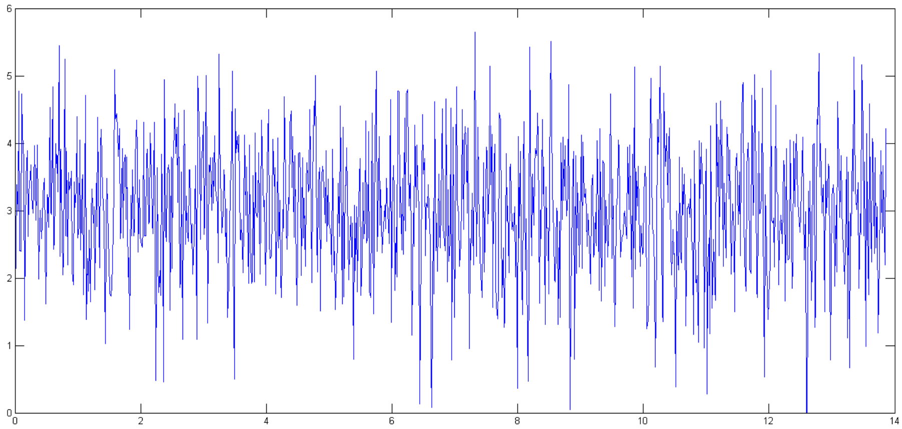
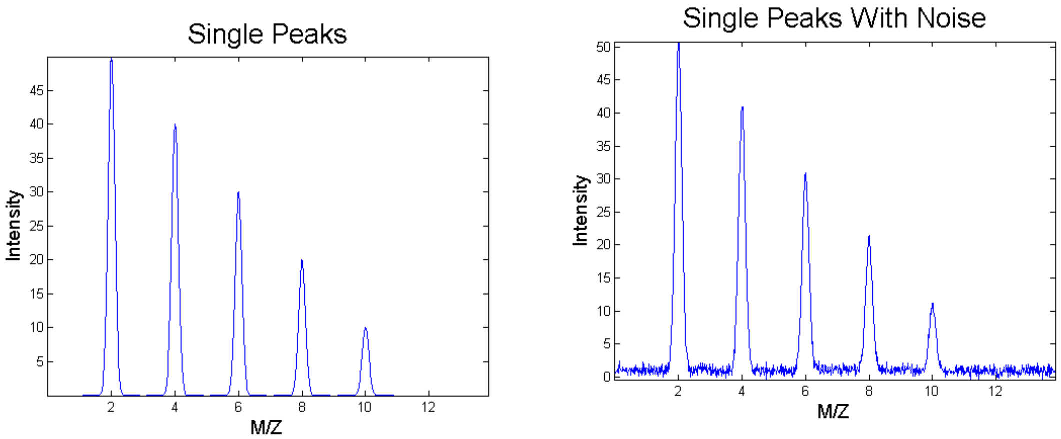


Figure 2.4: Model Gaussian noise.

Independent peak data

The simplest test of a peak picking algorithms performance is on standalone peaks. To test this, data with five standalone peaks were generated. Peak heights vary from 50 counts down to 10 counts. The centre of each peak is located on multiples of 2 along the mass axis. Each peak has a FWHM of 0.25 Da equating to a peak in a SIMS spectrum from around the 200Da. Figure 2.5 show these peaks before and after adding noise as described above. The noise in the Figure 2.5b has a mean of 1 and a standard deviation of 0.5.



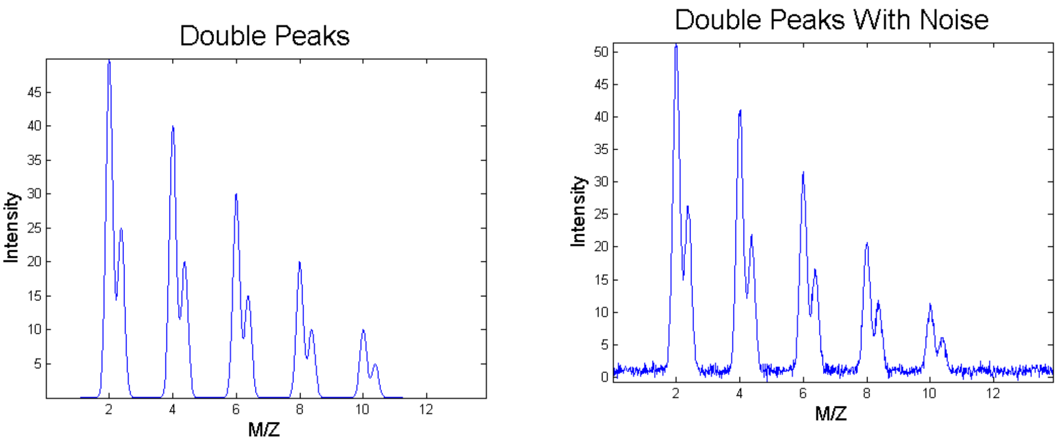
(a) Single peaks no noise.

(b) Single peaks with noise of mean 1 and standard deviation 0.5.

Figure 2.5: Single peaks without noise (2.5a) and with noise (2.5b).

Double peak data

More complex peaks are often found in SIMS data. To model this a ‘double peak’ set was created. The first peak in each set is again centred on the integer units. However, a second peak is added to each peak from the single peak data. The second peak is located 1.5 FWHM away and half as intense as the single peak data. Figure 2.6 shows an example of this with and without noise added.



(a) Double peaks no noise.

(b) Double peaks with noise of mean 1 and standard deviation 0.5.

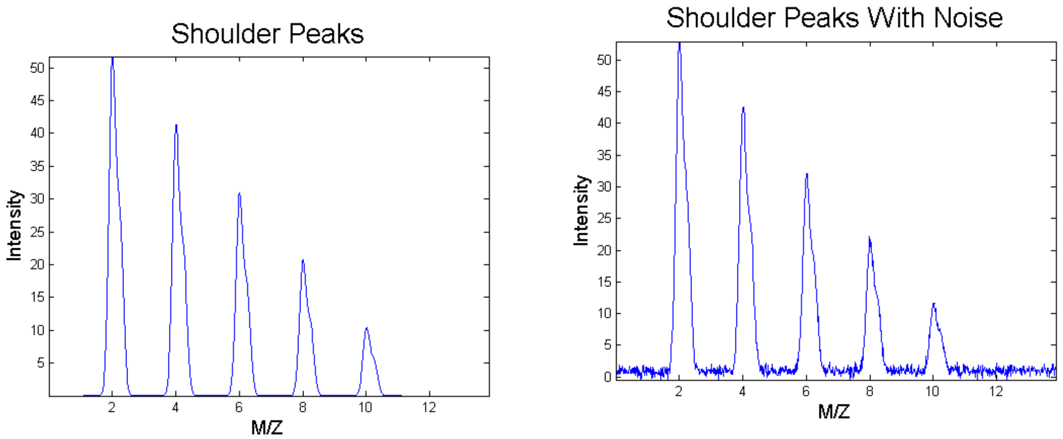
Figure 2.6: Double peaks without noise (2.6a) and with noise (2.6b).

Shoulder peaks

In high resolution SIMS data, peaks can often contain ‘shoulder peaks’ as discussed above. Data simulating shoulder peaks was generated (Figure 2.7). Additional peaks are again added to the single peak data. The second peak in each set are this time only one FWHM away from the parent peak and half as intense. This gives an obvious deviation from a standard peak shape but does not give a second maximum in the set or a clear trough between peaks.

Converging peaks

Data that encompasses double peaks and shoulder peaks can be seen in Figure 2.8. Here there are six sets of peaks. The first peak in each case is 50 counts high and is

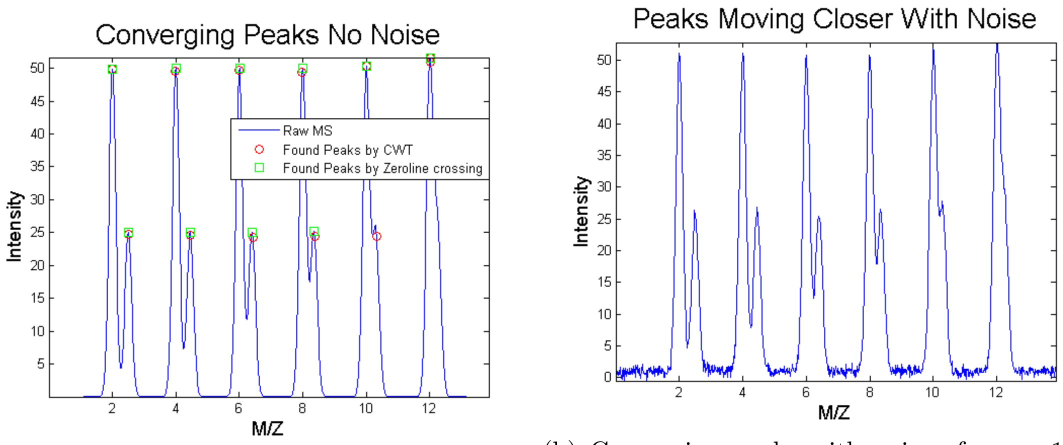


(a) Shoulder peaks no noise.

(b) Shoulder peaks with noise of mean 1 and standard deviation 0.5.

Figure 2.7: Shoulder peaks without noise (2.7a) and with noise (2.7b).

located on integer locations as before. The second peak in each set has an intensity of 25 counts. For each set the second peak "moves" 20% of the FWHM closer to the first peak, starting from two FWHM away i.e. starting at 0.5 Da away from the centre of the first peak and getting 0.05 closer in each set therefore 0.45, 0.4, 0.35, 0.3, and finishing 0.25 Da away in the last set. This means the last two peaks are located at 12amu and 12.25amu respectively.



(a) Converging peaks no noise.

(b) Converging peaks with noise of mean 1 and standard deviation 0.5.

Figure 2.8: Converging peaks without noise (2.8a) and with noise (2.8b).

Varying width peaks

To test the wavelet scale at which a maximum is observed across peaks of varying widths, data with peaks of varying widths were generated, Figure 2.9. The peaks follow sizes (number of data points in the peak) modelled on real Biotof data. Table 2.1 below describes rough estimates of peak sizes gathered from a Biotof spectrum. From these results the model data in Figure 2.9 were generated. The values of each peak can be seen in Table 2.2. This information can be used to seed the peak picking algorithm in terms of which wavelet scales to search for a given mass range.

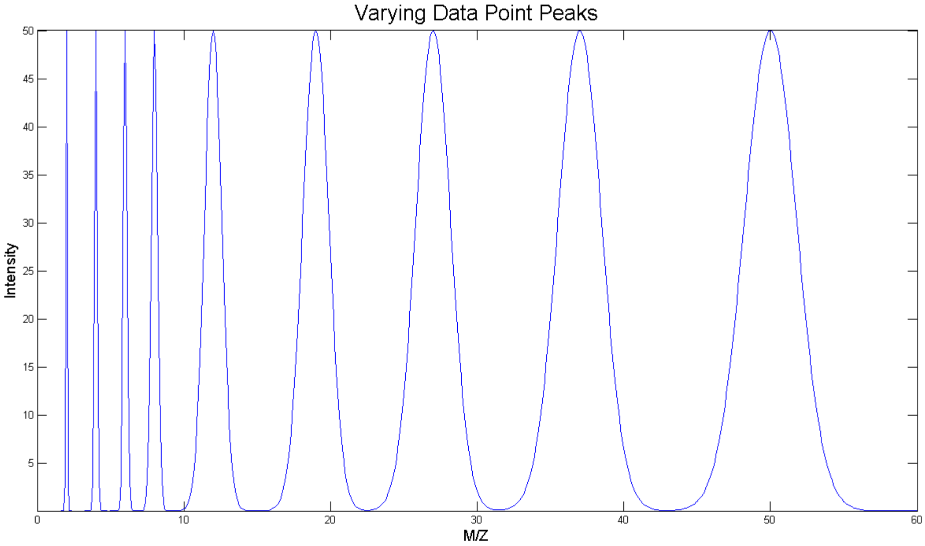


Figure 2.9: Peaks with increasing width.

2.5 Results

2.5.1 Wavelet methods comparison CWT *vs.* Zeroline crossing

Both the approaches adopted by Du *et al.* and Nyugen *et al.* incorporate finding peaks using multi scale wavelet transforms. Du *et al.*'s approach has been widely

Mass Range	Small Peak		Large Peak	
	Position	No. of data pts.	Position	No. of data pts.
0-50	26.03	18	41.07	27
50-100	57.73	19	57.14	44
100-200	136.5	30	134	50
200-300	236	30	217	49
300-400	332.9	50	354.7	80
400-500	447.9	70	441.6	70
500-600	562	70	542.9	90
600-700	640	72	663	80
700-800	744.5	80	781.8	80

Table 2.1: Rough estimates of number of data points in a selection of peak from various mass ranges.

FWHM	.1	.3	.5	.7	1.4	2.1	2.8	3.5	4.2
No. of pts.	3	7	11	17	44	49	67	83	99

Table 2.2: List of peak widths used to create data in Figure 2.9.

adopted and highly cited. This approach has also attained the highest results in comparisons against other peak picking routines for MALDI data [83]. Nyugen *et al.* claim to achieve a higher accuracy over the CWT method on SELDI-TOF data.

As described earlier Du *et al.* used the Mexican hat wavelet and formed ridge lines in the wavelet domain. Taking the maximum point on the ridgeline gave the centre of the peak and was proportional to the peaks parameters. Nyugen and colleagues' approach uses the first derivative Gaussian wavelet to locate peaks and created zerocrossing lines to find peak centres, they also used the Mexican hat wavelet to estimate peaks parameters.

To test these approaches thoroughly, both algorithms are stripped down and tested on the model data as described above. Results of these experiments are displayed as confusion matrices. Confusion matrices will take the form Figure 2.10:

Here true negatives will be set to zero as this category is simply the remainder of the unclassified mass channels in the spectrum and the value would be very large and nondescript for the purposes of the comparisons here. To illustrate; single peaks with no noise for both approaches gave results seen in the confusion matrices

		Prediction outcome		
		p	**n**	**total**
actual value	**p′**	True Positive	False Negative	P′
	n′	False Positive	True Negative	N′
	total	P	N	

Figure 2.10: Design of confusion matrix.

5	0
0	0

Table 2.3: Confusion matrix for Mexican hat on single peaks with no noise.

5	0
0	0

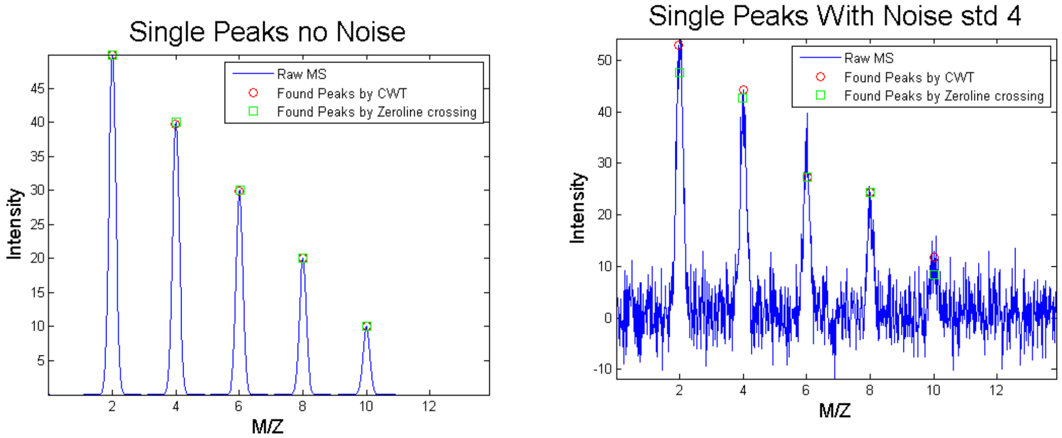
Table 2.4: Confusion matrix for zero-line crossing on single peaks with no noise.

in Tables 2.3 and 2.4 and Figure 2.11a. Here both approaches pick all 5 peaks with no false positives or false negatives.

Single peak with noise

With the addition of noise the problem of peak picking becomes more difficult but also more realistic. As the noise intensity increases the smallest peak's signal-to-noise ratio drops rapidly thus making it harder to decipher from the noise. As described in Section 2.4.1 noise is added with differing standard deviations ranging from 0.5 - 4 in 0.5 increments, thus there are 40 total peaks in the model data; 5 peaks with noise of 0.5 standard deviation added, 5 peaks with noise of 1 standard deviation etc..

These experiments were carried out by optimizing each of the two routines to be as accurate as possible while minimizing the number of false positives. Parameters were set constant across all levels of noise. Both approaches have two variables to



(a) Results for single peaks with no noise.

(b) Results for single peaks with noise standard deviation 4.

Figure 2.11: Results of Mexican hat and zeroline crossing for single peaks without noise 2.11a and with noise 2.11b.

40	0
0	0

Table 2.5: Confusion matrix for CWT on single peaks with eight levels of noise.

40	0
0	0

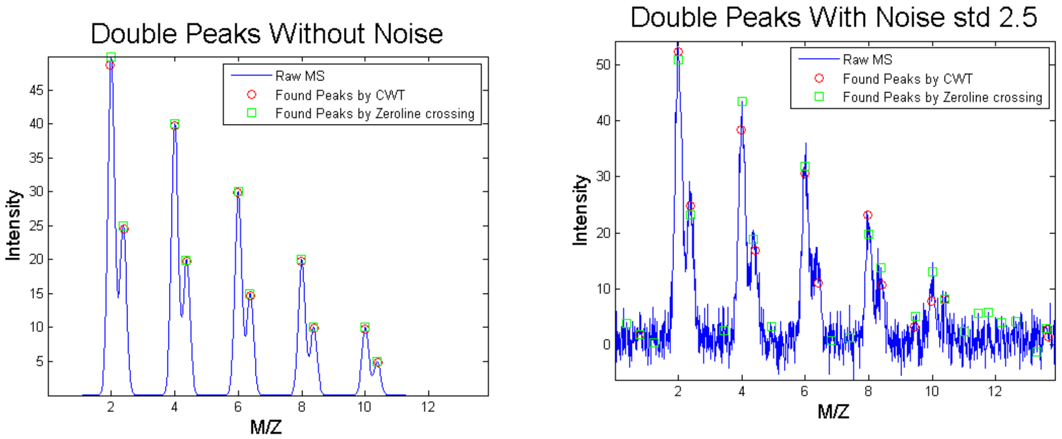
Table 2.6: Confusion matrix for zeroline crossing on single peaks with eight levels of noise.

set: the number of points along the ridge/zerocrossing line to be considered a peak and the number of scales to be used. However, the CWT approach has the advantage that a minimum threshold can be set to remove points below a certain threshold in the wavelet domain, thus these cannot be considered maxima and included in a ridge line. This helps reduce false positives. Looking at the confusion matrices (Tables 2.5 and 2.6), both approaches accurately pick all 40 peaks without picking any false positives.

Double peaks with noise

In SIMS data, a spectrum containing all single peaks is not observed in general applications; this will never be observed if looking at biological data using current instrumentation. Realistically peaks are found to overlap. These overlapping peaks again increase the difficulty of the problem of automated peak picking. Peak picking routines have to be accurate enough to discriminate each peak, not pick two peaks as a single peak, while limiting the number of false positive from noise. The double peak model data gives a more realistic approximation to real data than the simple single peak model data. With the addition of the model noise there are 80 total peaks in these data. Again, parameters were tuned, and kept constant, for sensitivity while limiting the number of false positives.

Examination of the confusion matrices (Tables 2.7 and 2.8) indicates that both routines did identify correctly most of the peaks. The CWT approach scoring slightly higher here. In the false positive category the CWT well outperforms the zero-line crossing. This shows that the balance between sensitivity and specificity is favourable in the CWT approach. The number of false positives is less than half of the zeroline crossing method. Some discrepancy can be seen between the picked centres of the peaks in Figure 2.12b, this will be discussed further in Section 2.5.1.



(a) Results for double peaks with no noise.

(b) Results for double peaks with noise standard deviation 2.5.

Figure 2.12: Results of CWT and zeroline crossing for double peaks without noise 2.12a and with noise 2.12b.

78	2
50	0

Table 2.7: Confusion matrix for CWT on double peaks with eight levels of noise.

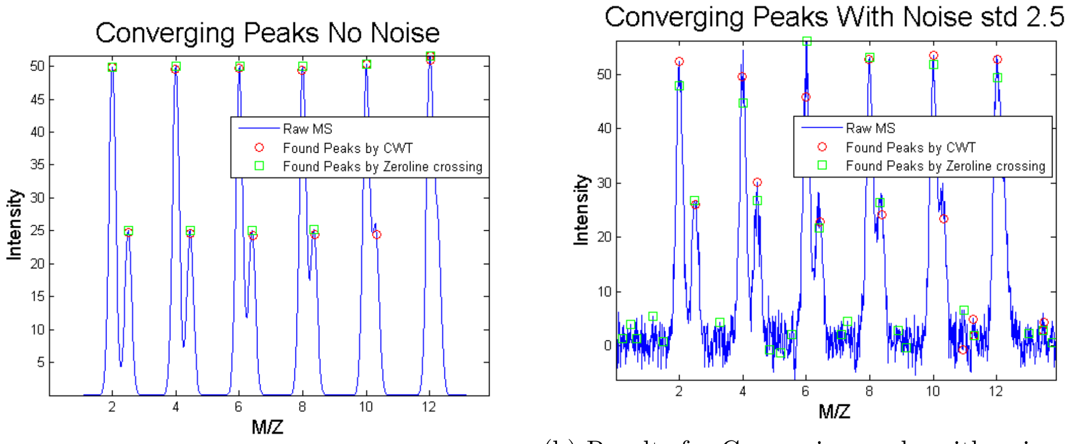
76	4
115	0

Table 2.8: Confusion matrix for zeroline crossing on double peaks with eight levels of noise.

Converging peaks with noise

The converging peaks data were designed to simulate a range of peaks that occur in real SIMS data. Peak overlap varies widely in real data and an analytic approach to investigating this phenomenon was desired. Similarly to previous analysis, model noise was added making a total of 96 peaks with varying overlap.

As with the double peak data the CWT method outperforms the zeroline crossing method. Here the CWT picked 90 peaks correctly, with only 33 false positives. This is compared to 79 correct and 113 false positives for the zeroline crossing. A clear difference between the ability of each approach to remain sensitive while excluding unwanted noise is observed. The zeroline crossing method cannot pick the small features and remain sensitive to noise exclusion. This is discussed more in the next section.



(a) Results for Converging peaks with no noise.

(b) Results for Converging peaks with noise standard deviation 2.5.

Figure 2.13: Results of CWT and zeroline crossing for Converging peaks without noise 2.13a and with noise 2.13b.

90	6
33	0

Table 2.9: Confusion matrix for CWT on Converging peaks with eight levels of noise.

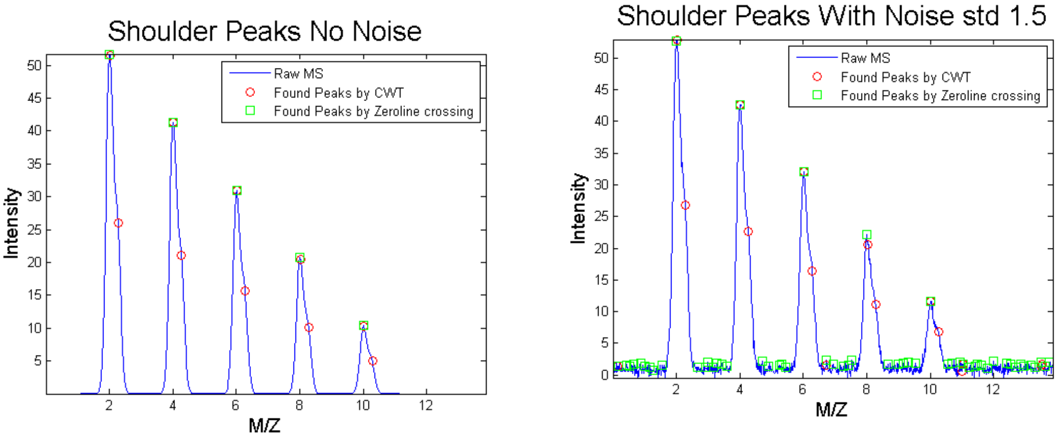
79	17
113	0

Table 2.10: Confusion matrix for zeroline crossing on Converging peaks with eight levels of noise.

Shoulder peaks

Shoulder peaks are a common occurrence in real SIMS data. This is especially true of high resolution machines. The ability to find these peaks can give descriptive information about the sample. In Chapter 3 a use case for finding such peaks is described.

Observing Figure 2.14, similar to the converging peaks example the zeroline crossing method was unable to classify the shoulder peaks, even when set to maximum sensitivity. The CWT approach however had little trouble discovering them. As can be seen in Figure 2.14 when some noise is added the CWT method does not classify many false positives even though it is very sensitive to the shoulder peaks.



(a) Results for Shoulder peaks with no noise.

(b) Results for Shoulder peaks with noise standard deviation 1.5.

Figure 2.14: Results of CWT and zeroline crossing for Shoulder peaks without noise 2.14a and with noise 2.14b.

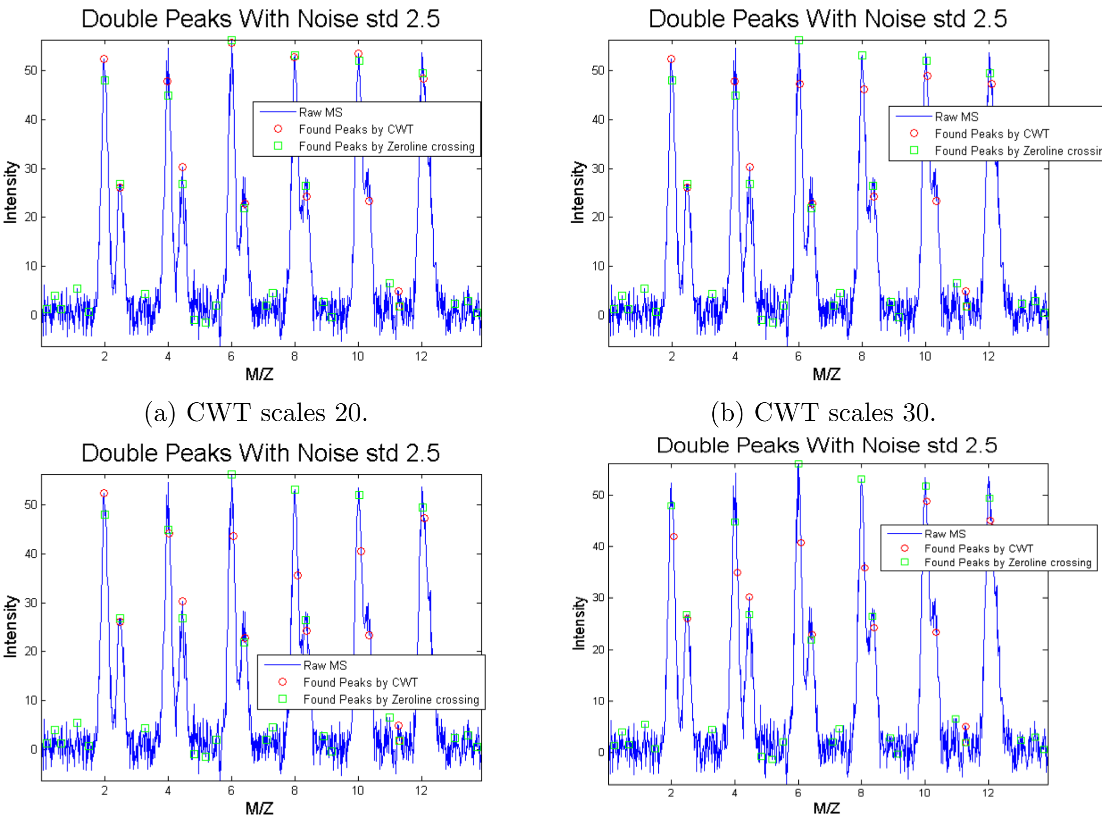
Identification of centre of the peak

The accurate location of peak centres is one of the goals of this type of analysis. Above favourable aspects of the CWT approach have been experimentally discovered. One of the pitfalls of the CWT approach can be the accuracy of finding peak centres across large scale differences. For single peaks it performs very well. But real world spectral analysis where there is noise and peak overlap this can cause discrepancies in the peak identification. To simulate large scale differences the number of input scales to the CWT approach was varied from low, therefore accurate at small scales, to very high, accurate at very large scales. The results of this on double peak data can be seen in Figure 2.15.

Starting with Figure 2.15a the input number of scales, 20, matches well with the data. In Figure 2.15a-2.15d the picking of the location of the large peaks gets pulled toward the centre of the combination of the large and small second peak. This is a clear result of the fact that at larger scales the maximum point on the ridge line is now at the wavelet coefficient that includes a combination of both peaks. The zeroline approach results do not change in the figures due to scales not being varied. As was described in Section 2.3 a combination of both methods is adopted to avoid these effects. The CWT approach is highly accurate across large scales while remaining resilient to noise distortion. However, in cases like this a second vote on the estimated centre of the peak could avoid misclassification. By keeping the zeroline method at a sensitive level these zerocrossing lines can be used instead of the CWT centres.

Varying width peaks

Figure 2.16a shows the model data described in Section 2.4.1. These data were generated to test the scales at which the Mexican hat wavelet coefficients resonate with peaks of different widths. Figure 2.16b shows the results of applying the Mexican hat wavelet across 60 scales to the data in Figure 2.16a. Table 2.11 describes the results of this experiment. From this table a guide to the input scale across different mass ranges can be generated. Thus, allowing the peak picking method to be sensitive to low intensity or sharp peaks while remaining broad enough to categorise large or broad peaks at different mass ranges.



(a) CWT scales 20.

(b) CWT scales 30.

(c) CWT scales 40.

(d) CWT scales 50.

Figure 2.15: Results of varying the CWT scale from 20-50.

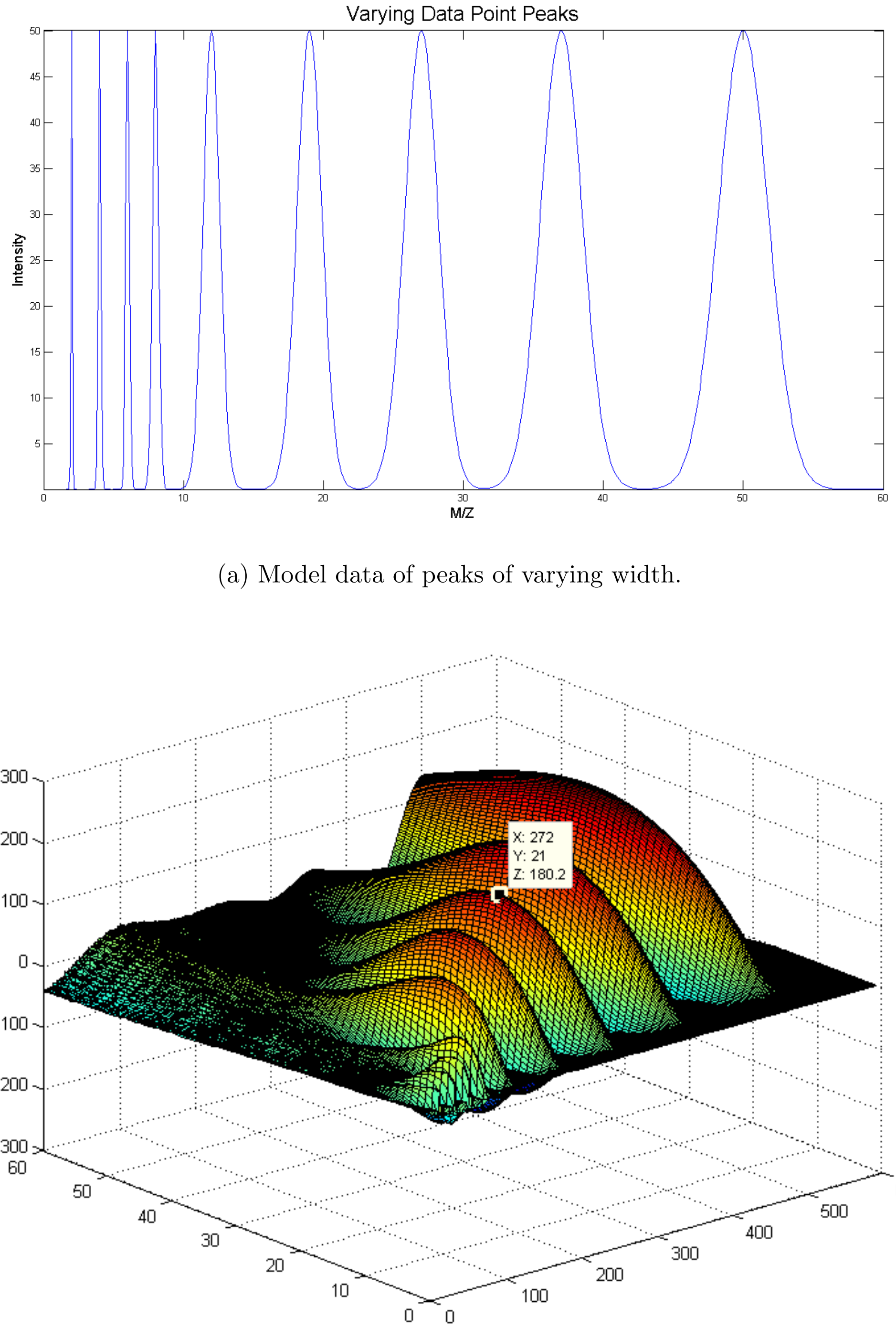
FWHM	.1	.3	.5	.7	1.4	2.1	2.8	3.5	4.2
Maximum wavelet scale	2	2	3	5	11	17	21	26	33

Table 2.11: List of the FWHM and number of data points in the peaks in Figure 2.16.

2.5.2 Full Routine Peak Picking

As was noted by [13] the number of peaks found per spectrum is not, by itself, an adequate measure of the quality of a peak finding algorithm. It is important to ascertain if the peaks being found by the algorithm correspond to real phenomena in the spectra. To test the efficacy of all the algorithms described above against the approach adopted here, real data acquired on the Biotof were used. Here peak picking was applied to a spectrum to test the accuracy and adaptability of each of the approaches studied in terms of real world data.

A selection of mass ranges were extracted to visualise the performance of each



(a) Model data of peaks of varying width.

(b) 3d representation of applying the Mexican hat CWT to the data in 2.16a.

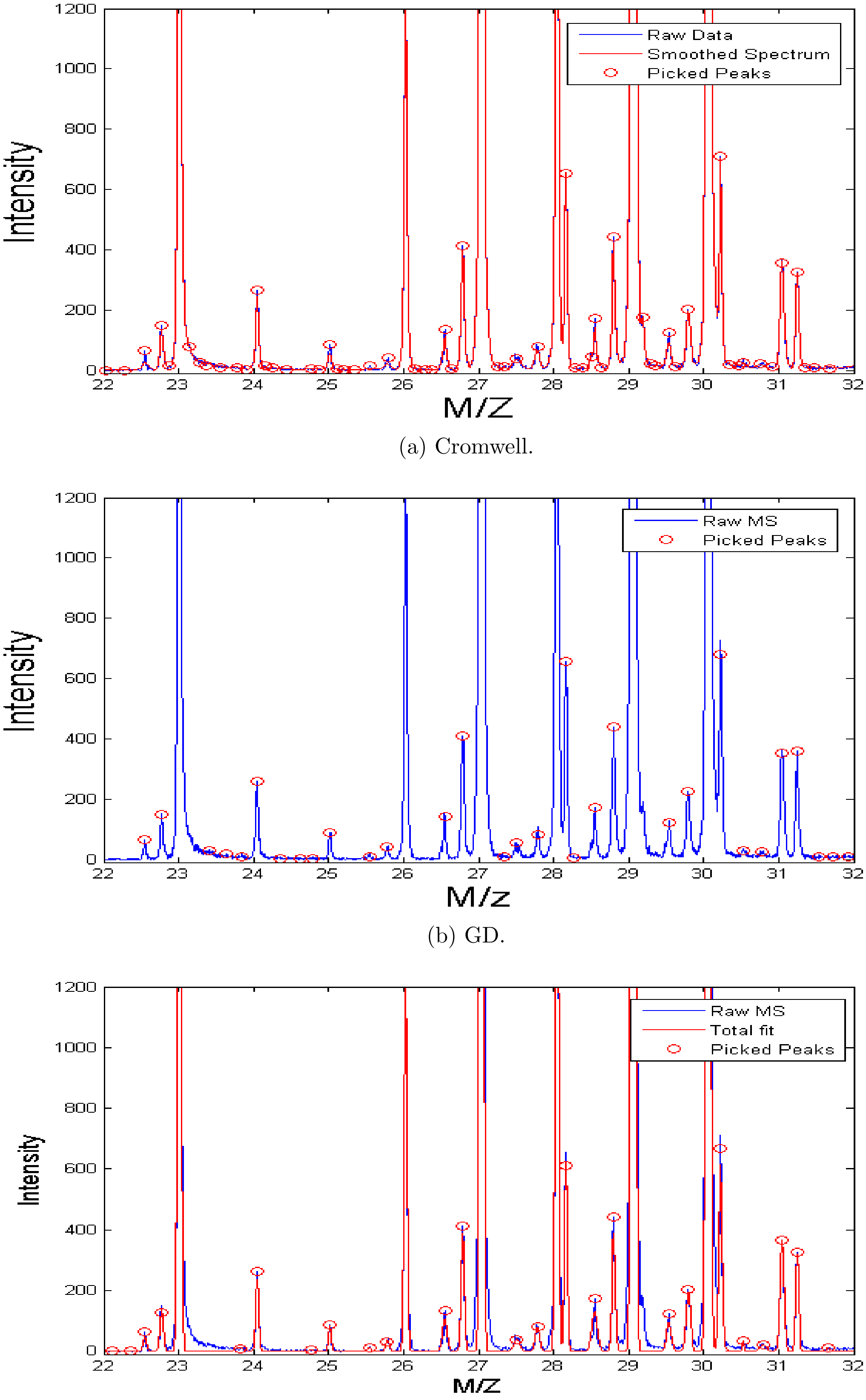
Figure 2.16: Peaks of varying width 2.16a. 2.16b 3D representation of the wavelet coefficients applied to 2.16a.

of the approaches. A low, middle and high mass range are used to best display how the peak picking routines perform across a full spectrum. Figure 2.17 shows the results of peak picking from the Cromwell method, the GDWavelet method (zeroline crossing) and the approach adopted here based on the CWT method for a low mass range of 22 Da - 32 Da. All routines had their particular parameters tuned to maximise the number of true positives while limiting false positives.

From Figure 2.17 it is clear that all three methods perform well on this mass range. All three methods pick all the visibly observable peaks. However, Cromwell (Figure 2.17a) picks a larger proportion of false positive than the two wavelet approaches. Comparing GDWavelet (Figure 2.17b) and the new routine (Figure 2.17c) there is little difference except the GDWavlet picks slightly more false positives for this mass range.

Figure 2.18 shows the results of three approaches applied to the 104-114.5 Da mass range. Here some overlapping and shoulder peaks are observed. Again all three approaches pick the clear peaks. There are three shoulder peaks in this set located at 111.2, 113.2 and 114.2 Da. GDWavelet the zeroline crossing approach does not select these as peaks, Figure 2.18b. Cromwell selects the two more pronounced peaks at 113.2 and 114.2, Figure 2.18a. Only the new approach selects the subtle shoulder peak at 111.2 Da. Also both the Cromwell and GDWavelet approach select more erroneous small peaks than the new approach.

Continuing up the mass range in the SIMS spectrum the signal intensity drops. Figure 2.19 shows the results of the three methods for the mass range 503-513 Da. One of the limitations of the Cromwell approach becomes very evident at this mass range (Figure 2.19a). Here the smoothing that was applied to maintain specificity at lower scales over smooths at higher scales. Thus a lot of peaks are not selected in this mass range where signal-to-noise is very low. Both wavelet approaches continue to select peaks well. However, GDWavelet selects more false positives again for this mass range (Figure 2.19b). The new approach clearly out performing the other two.

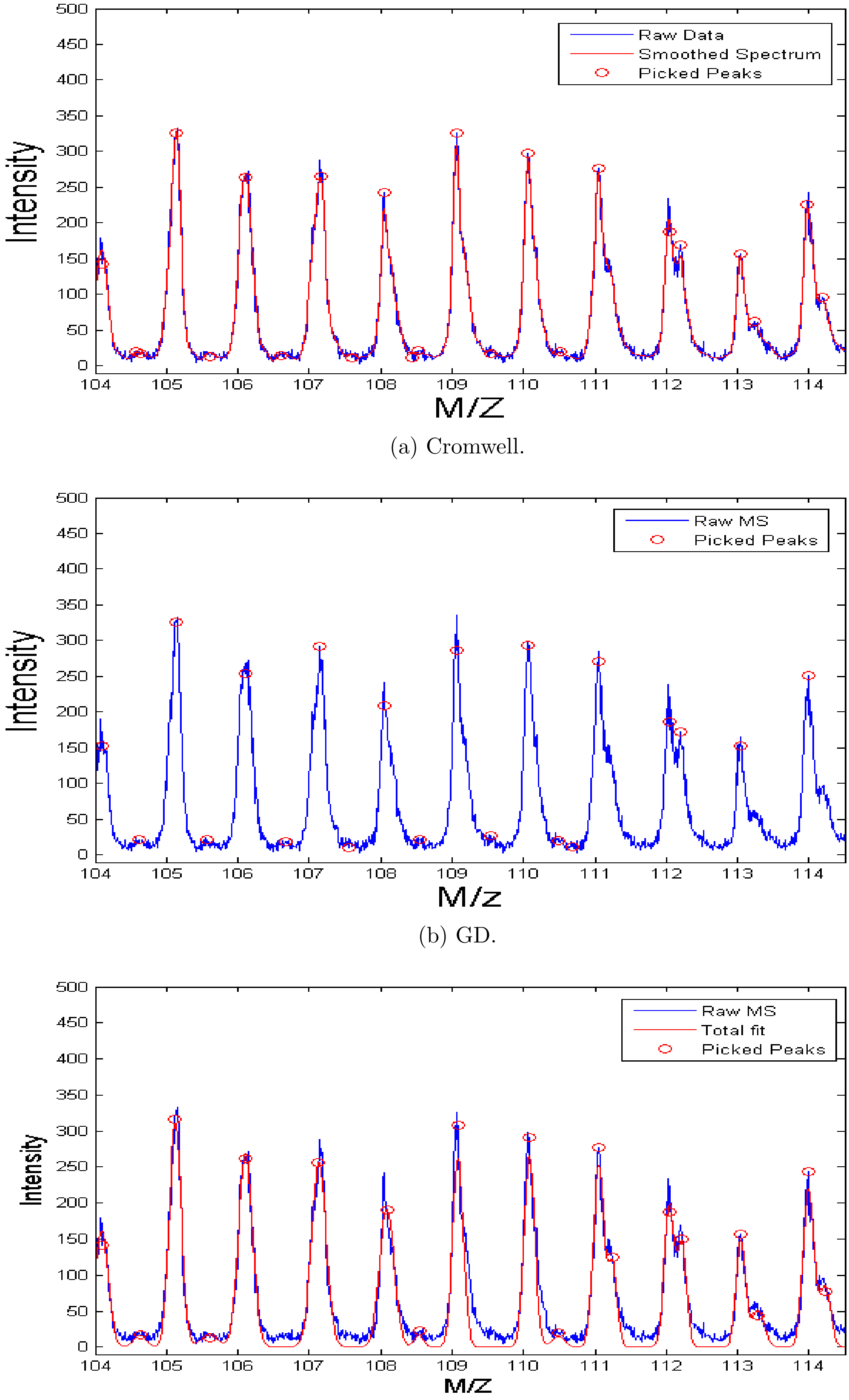


(a) Cromwell.

(b) GD.

(c) Approach outlined in this thesis.

Figure 2.17: Results of Cromwell 2.17a, GDWavelet 2.17b and new routine 2.17c for a section of a Biotof spectrum, mass range 22-23 Da.

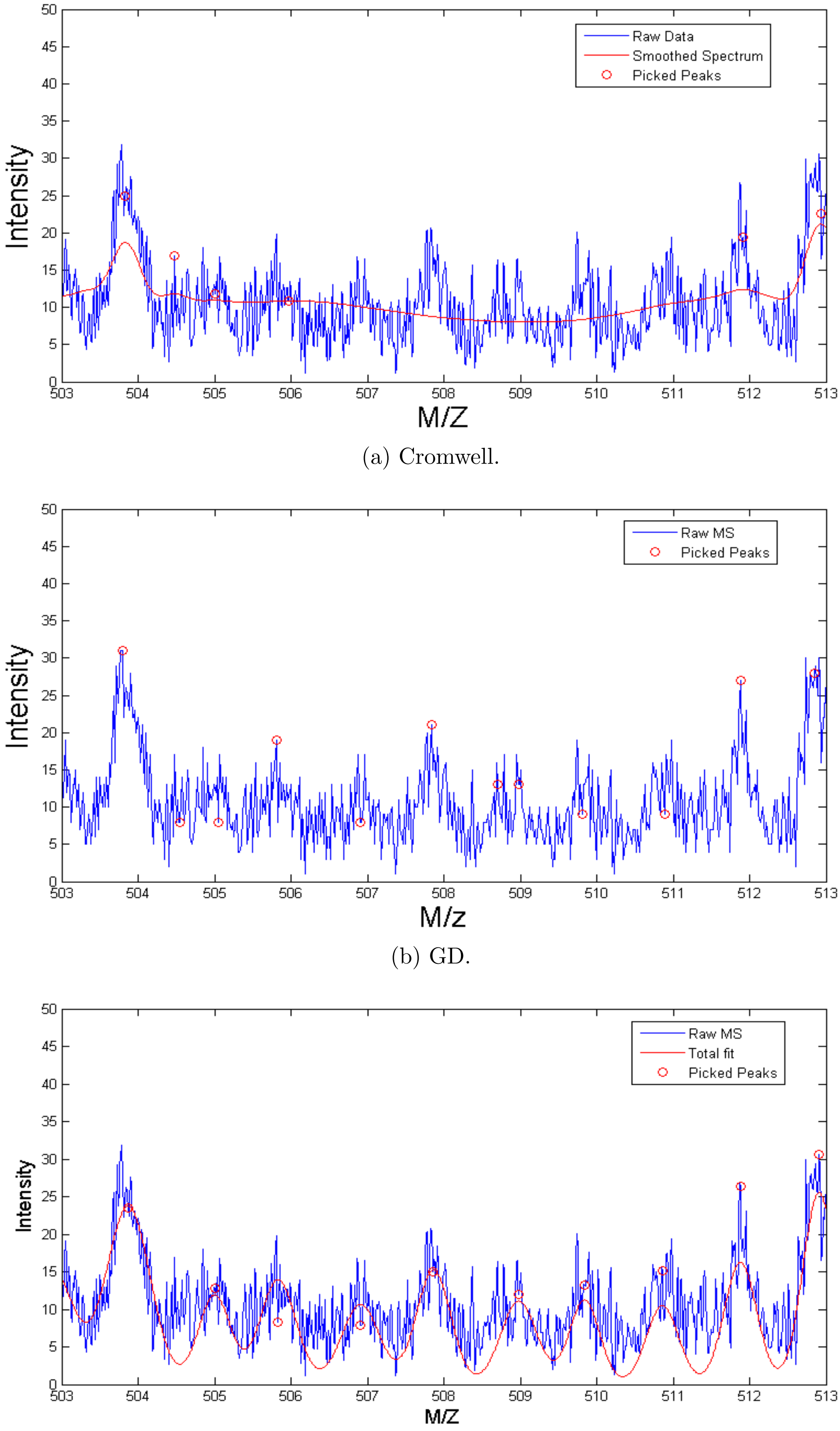


(a) Cromwell.

(b) GD.

(c) Approach outlined in this thesis.

Figure 2.18: Results of Cromwell 2.18a, GDWavelet 2.18b and new routine 2.18c for a section of a Biotof spectrum, mass range 104-114 Da.



(a) Cromwell.

(b) GD.

(c) Approach outlined in this thesis.

Figure 2.19: Results of Cromwell 2.19a, GDWavelet 2.19b and new routine 2.19c for a section of a Biotof spectrum, mass range 552-568 Da.

2.6 Discussion

In Chapter 3 peak picking is applied as an input into a compression routine for ToF-SIMS images. Depending on the desired needs of the analyst this peak picking routine can be tailored accordingly. If largest compression is desirable then discovery of shoulder peaks and low intensity peaks may not be a necessity. However, if retention of data is the desired outcome of peak picking the routine then it can be tuned to gather maximum amounts of information. This may include noise but will give a result closer to maximum signal retained.

Advantages to this approach over previous approaches are the fact that it can easily be tuned for SIMS data. High accuracy across a whole spectrum can be achieved. Also, the peak fitting routine can compensate for errors in the peak parameter estimation and the peak centre estimation. As was noted these methods can fall prey to overlap distortions. This is especially true in the case of parameter estimation. The fitting routine can compensate for this.

As was noted to reduce the computation time, the fitting routine is passed sections of the spectrum to fit, 2 Da for example. This can cause errors in the fitting due to the routine trying to minimize the errors on a fit within this window and can fit a peak much too large if there is overlap with another peak just outside the window. To compensate for this, bounds need to be set in a reasonable on the fitting routine to limit the scope for fitting.

A fast version of the routine has also been produced which excludes the calculation of estimation of parameters and peak fitting. This significantly reduces the runtime. This type of analysis may be slightly less accurate due to not peak fitting, especially in terms of shoulder peaks. However, if the location of peaks is all that is desired this gives a quick approach without losing much accuracy.

2.7 Conclusions

In this chapter peak picking routines were applied to ToF-SIMS data. An investigation into the most applicable method to apply to SIMS data was carried out. The results of this indicate that an amalgamation of the CWT method and the zeroline crossing gives the best results for diverse sets of SIMS data. An extension to these and previous methods is the fitting of a generalized distribution to increase accuracy and to separate overlapping signal.

Accuracy on real data of was achieved across diverse SIMS data. Peak separation was deemed a necessity as overlapping peaks distort the actual information that can be acquired by peak picking routines. Overlapping peaks and peak shoulders exemplify this. To extract an estimate of the actual information contained in these shoulders and overlapping peaks is essential to obtaining accurate representations of the chemistry contained within a mass spectrum. To do this; scripts were written to fit the model peak shape as best as possible to the convoluted raw data, thus allowing for the peaks parameters to be calculated rather than estimated as in a previous step and used in further calculations. Chapter 3 uses the peak picking routine as a pre-processing step in the analysis of SIMS Images.

Chapter 3

Compression of SIMS images

Contents

3.1 Introduction

The size of ToF-SIMS datasets has become an issue in recent times. Though computers are increasing in capability the data sets being acquired from SIMS instrumentation have also been increasing dramatically. Instrumentation with higher duty cycles produce images with more pixels more quickly [24]. 3D molecular imaging is now also possible [32], stacks of images that can easily reach tens of gigabytes are produced. Improved mass resolution and sensitivity also provide more discernible peaks [10]. The ability to create ‘tiled’ images of large surfaces [38] is also a source of massive files sizes. This chapter investigates methods of compression of ToF-SIMS data and proposes alternatives to the classical practices of binning data, where mass resolution is lost, and manual selection of mass ranges over peaks to image.

As stated previously the term ‘peak picking’ is used to describe the exercise of determining a ‘discrete spectrum from continuous data’. This is in contrast to arbitrary peak selection based on *a priori* knowledge of the sample [80] and/or by applying a simple intensity threshold to the data. Peak picking as described in Chapter 2 was used as an input to the approaches described here.

Peak picking is a method of locating peaks within a spectrum. Classically peak picking has been used to find peaks to generate peak lists. From this the area of each peak discovered can be used instead of its original or fitted distribution, thereby reducing the information to a single mass channel. There have been many approaches adopted to do this by the mass spectrometry community [17] [53]. In the previous chapter peak picking was applied to locate peaks, estimate their parameters and fit a generalised distribution. Here this information is applied to ToF-SIMS images to evaluate new and classical approaches for interpreting images.

Individual pixels, in general, do not contain enough information to accurately identify a peak’s position or distribution, due to lack of signal intensity within a given single pixel. However, by summing the signal from all the pixels in the image, this usually produces a spectrum with good signal-to-noise: the total ion spectrum. A peak list can be formulated from the total ion spectrum. Using this information, each of the constituent pixel spectra can be queried as to their values under a peak's distribution. However, if there are shoulders or overlapping peaks within the total ion spectrum, this complicates analysis. These spectral features must be 'de-convoluted' or separated to estimate realistic values to achieved more instructive compression than classical binning. By evaluating in this way, a highly sparse matrix for all spectra contained in an image can be compiled. In-house testing indicates that this compression method can give a compression of over 95%, e.g. assuming four discernible peaks per Da and a mass range of $1 - 1000$ Da, with $100,000$ mass channels, a compression of 96% is achieved. This allows for much faster statistical analysis to be performed without the need for massive RAM requirements and opens the door for more computationally intensive methods of data processing to be adopted, e.g. MVA.

Here three new methods of processing images are proposed. These methods are explained in Section 3.3 and the results of these methods are analysed using model and real data in Section 3.4 and [50].

3.1.1 Imaging a mass range

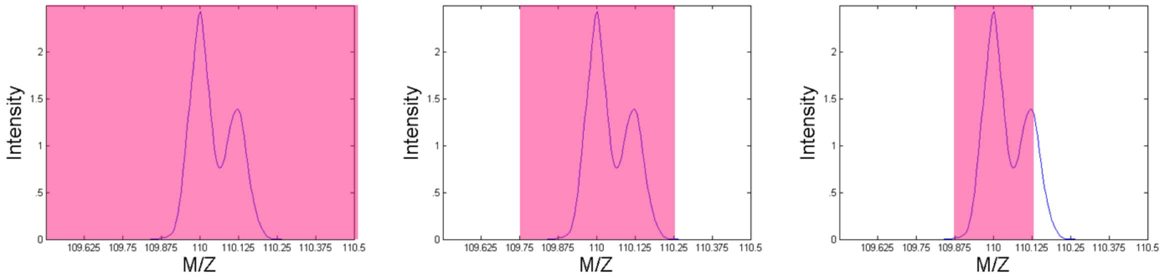
When imaging a peak, the analyst can manually sum across a portion of a peak to generate an image. This can be representative of the signal's distribution throughout the image. However, this is very dependent on the mass limits manually chosen by the user; an inherently subjective approach. This is especially true if peaks are overlapping. If the chosen bound is set towards the neighbouring, overlapping peak, the more likely it is to erroneously contain signal from that neighbouring peak. However, by moving the bound away from the neighbouring peak some of the true distribution will be lost from the image resulting in a reduction in contrast. The proposed techniques outlined here aim to limit the false signal that can be assigned to a peak by manual peak selection arising in an image while maintaining maximum signal levels and thus contrast. Methods described below will also give automated, reproducible methods for all peaks/images. This has the advantage of averting the subjective and near irreproducible nature of manual selection and can be applied repeatedly to different peaks/images.

3.1.2 Binning

Since manual imaging of peaks is generally not performed for all peaks in the spectrum, a common approach adopted to perform exploratory analysis is binning. When data are binned, either to nominal mass or some other arbitrary unit, the distribution of the original signal is lost and so the true peak position, the centre of that distribution is also lost. The binning method simply sums mass channels together in a regular fashion. The advantage of this is that signal is summed together, therefore the amount of data are reduced and signal distributions are also condensed. However, the latter is also one of the disadvantages of binning data. Signal distributions are not considered and as a result can be spread across two or more bins. This implies that signal arising from the same distribution can be separated and considered differently as far as multivariate methods are concerned. Conversely distributions that are independent of each other can be put into the same bin. This gives an unnecessary mixing of the data.

With recent advances in instrumentation mass resolution has increased, however mass resolution is lost when data are binned resulting in an effect contrary to the developments in instrumentation. As has been previously noted [36], different pretreatments of images can have a large effect on MVA techniques and some binning routines were noted as having a detrimental effect on certain types of MVA.

Figure 3.1 demonstrate some general binning routines and the effect they have when applied to some arbitrary peaks. In Figures 3.1a-3.1c the pink region represents the size of the binning window, the convention of binning around the nominal mass unit was followed for these examples. It is apparent that in Figure 3.1a the two peaks shown were binned into the same bin. The bin here is 1 Da wide placed about the nominal mass unit. This style of binning is common and as is illustrated vast amounts of information are lost by this approach. By reducing the size of the bin in



(a) Bin width of 1 Da.

(b) Bin width of 0.5 Da.

(c) Bin width of 0.25 Da.

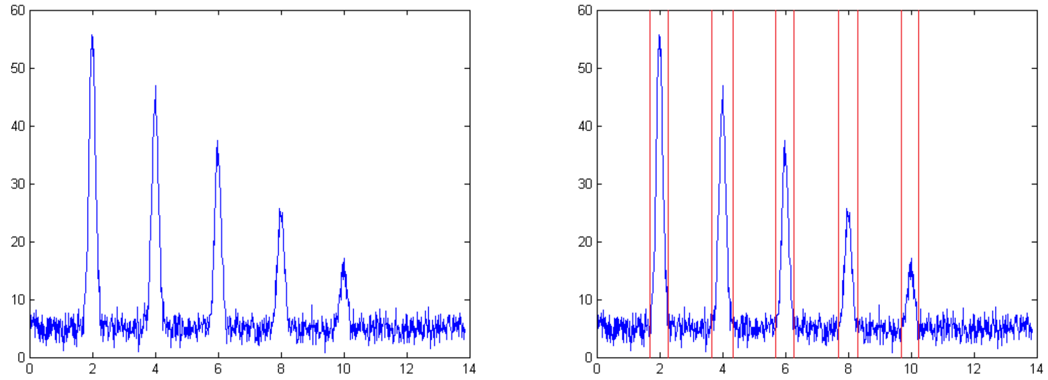
Figure 3.1: Visual representation of common bin sizes over peaks.

Figures 3.1b and 3.1c the amount of data summed to a bin is smaller, thus allowing for better resolution of the peaks. However, in Figure 3.1c the bin size still does not fit well with the data and some of the second peak will be put within a different bin. Also, the smaller the bin size the smaller the compression thus the greater the size of the processed data. In Section 3.3 three new methods are proposed as alternatives to binning for processing image data.

3.1.3 Problem statement

The aim of this chapter is to illustrate alternatives to manual peak selection or binning of data. all methods below aim to compress image data by finding peaks (peak picking) and compressing these peaks into a single mass channel at the peaks estimated centred. Figure 3.2a depicts the total ion spectrum of simple model data. As opposed to simply binning the data here the processes estimate the peaks position and compress the peaks using some bounds (Figure 3.2b). By approaching the data in such a way the compression acts as a noise filter by excluding noise from the compression.

For such simple data this seems trivial. However, SIMS spectra are much more complex. The problem attacked in this chapter is how to deal with more complex features in an image. The processes outlined below are all methods to combat complex SIMS data. SIMS spectra include much more features such as overlapping peaks and shoulder peaks. All the routines outlined below perform equally well on



(a) A model spectrum.

(b) A model spectrum indicating peak bounds.

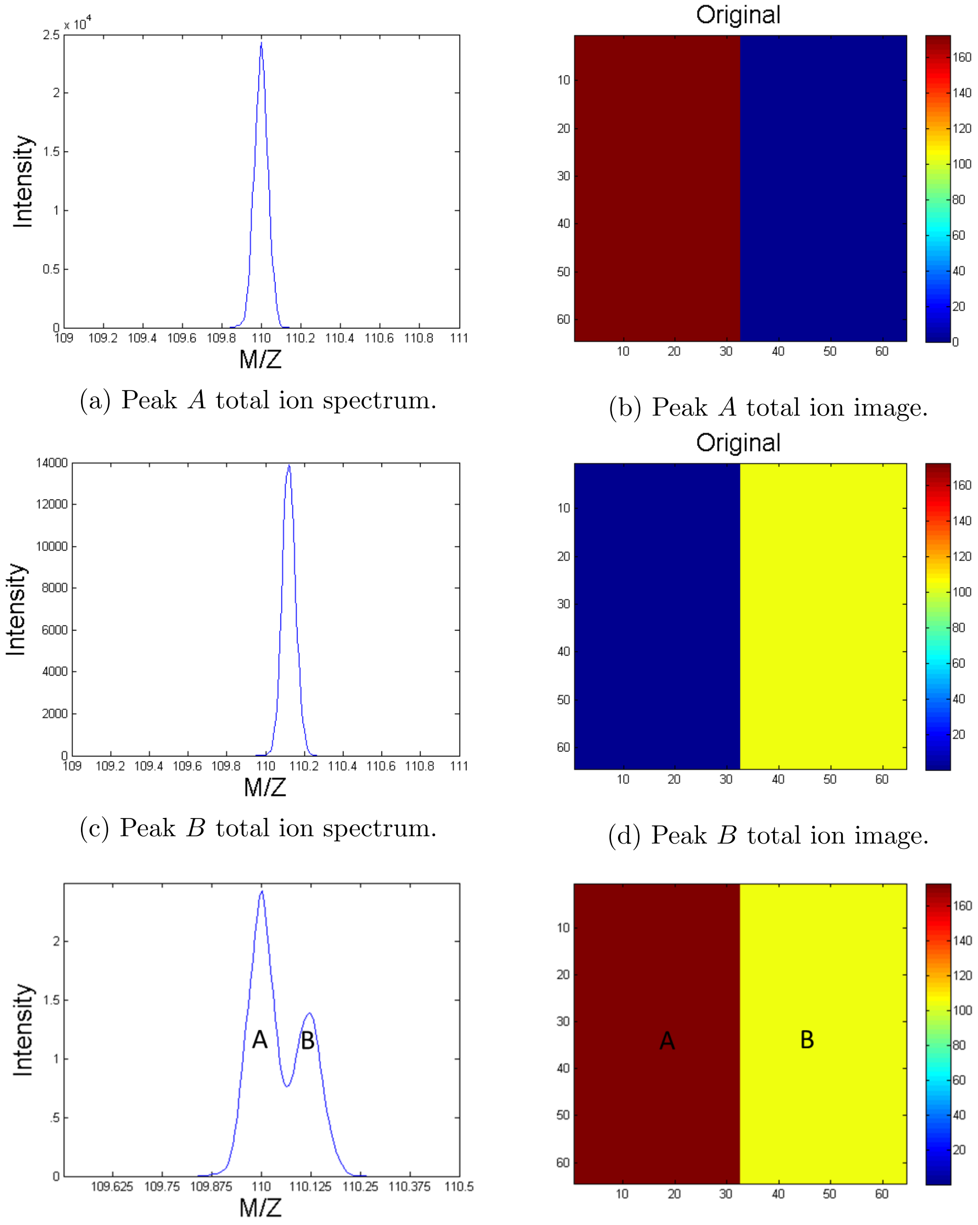
Figure 3.2: Model data with 3.2a and without 3.2b some defined peak bounds.

single isolated peaks, such as those in Figure 3.2, which can easily be compressed. To compare the suggested methods of compression some model data were created for ease of direct comparison. Real instrument data were also acquired using two specifically chosen samples that exhibited the desired characteristics for testing and will be discussed in Section 3.4.

3.2 Data

3.2.1 Model data

A simple data set was simulated in order to produce two distributions of signal overlapping in the total ion spectrum, but spatially separated in the total ion image. Here the Box-Muller transform [6] was utilised to generate a randomised collection of Gaussian peaks of equal intensity, with Gaussian distribution on the mass scale. This transform simply generates a list of normally distributed numbers from a list of random numbers, these then are used as the location of the signal intensity within the pixels of the generated images. The procedure was repeated for two Gaussian peaks of different intensity and which exhibited an overlap. These separate distributions were then arranged in an image such that all pixels on the left of the image had signal arising from peak A (left peak) and those from the second, lower intensity peak B (right peak) had signal to the right of the image as shown in Figure 3.3a - 3.3f. Figures 3.3e and 3.3f show these data combined in both the total ion spectrum and the total ion image. Following the same approaches outlined for the two peak model data above, a three peak model data set was also generated. The results of which can be seen in Figures 3.4.



(a) Peak A total ion spectrum.

(b) Peak A total ion image.

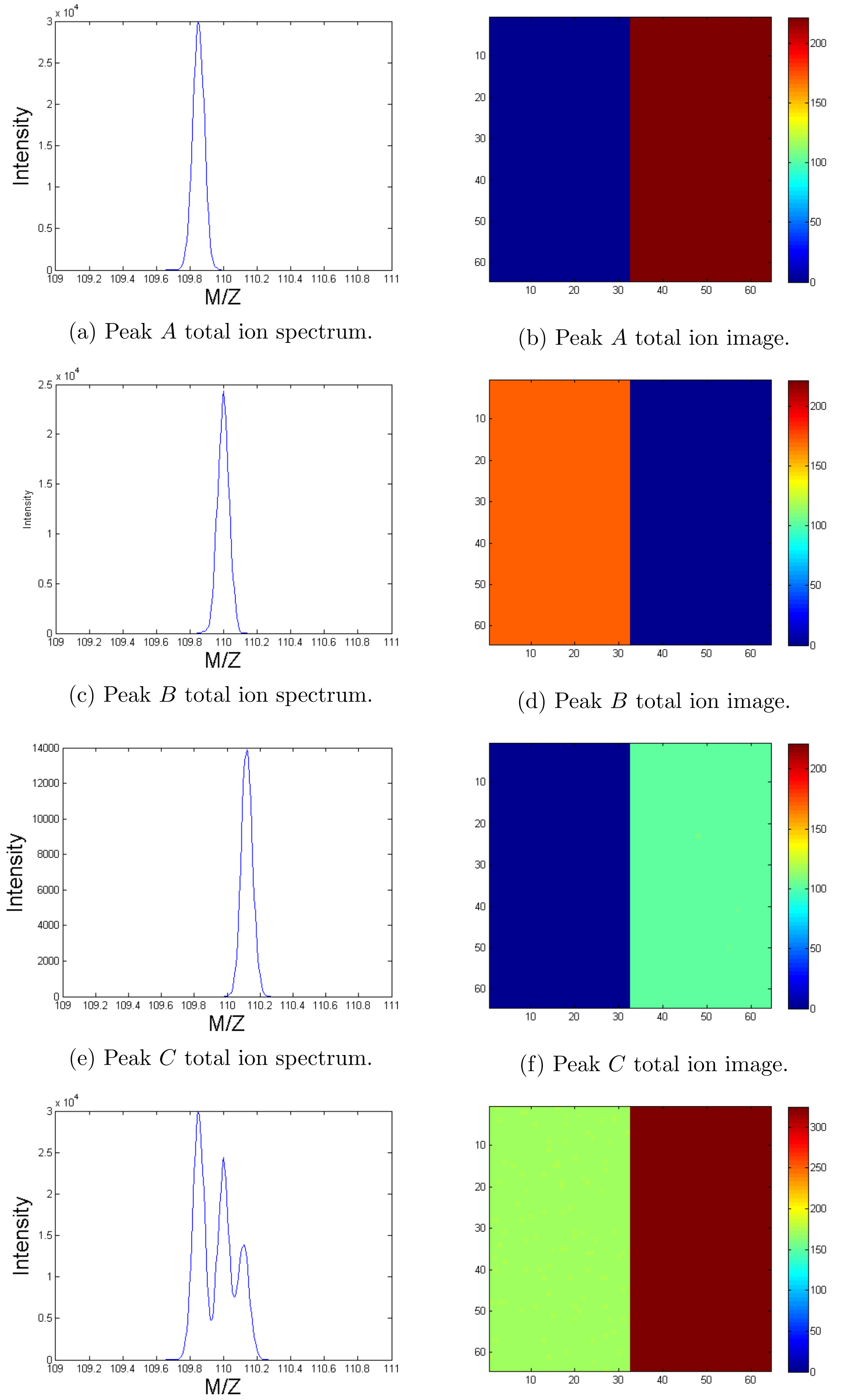
(c) Peak B total ion spectrum.

(d) Peak B total ion image.

(e) Total ion spectrum of A and B combined.

(f) Total ion image of A and B combined.

Figure 3.3: Images of the two peak model data.



(a) Peak A total ion spectrum.

(b) Peak A total ion image.

(c) Peak B total ion spectrum.

(d) Peak B total ion image.

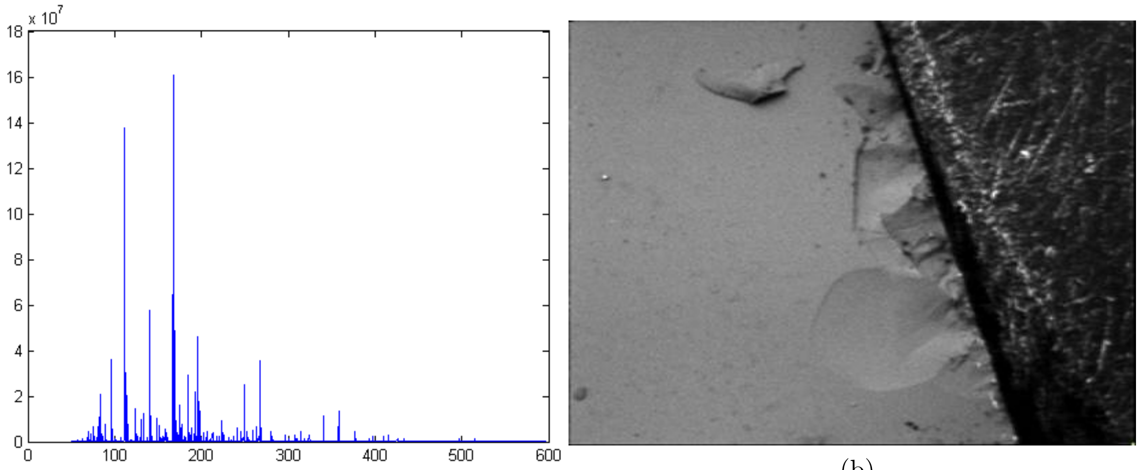
(e) Peak C total ion spectrum.

(f) Peak C total ion image.

(g) Total ion spectrum of peaks A, B and C.

(h) Total ion image of peaks A, B and C.

Figure 3.4: Images for the three peak model data.



(a)

(b)

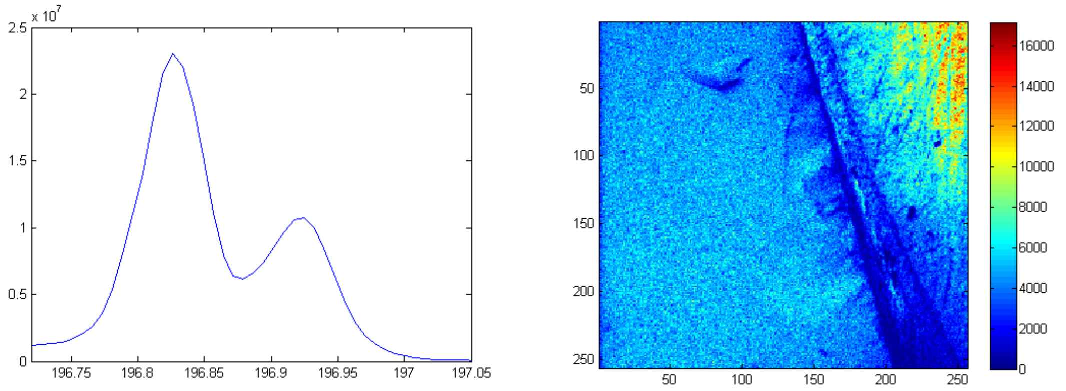
Figure 3.5: Total ion spectrum 3.5a and total ion image 3.5b of a sample containing regions of silicon oxide and tantalum oxide.

Rationale

The motivation for creating model data in such a manner was to choose a simple example that clearly defines the strengths and weaknesses of each approach of data analysis and to give a simple visual metric, as well as a percentile score, to highlight the potential benefits of the methods. These model data are clearly an over simplification but is used to exemplify the proposed approaches. Since peaks generated from the J105 are observed to be generally Gaussian these model data give a realistic, if simplistic, estimation of actual data.

3.2.2 Acquired data

Considering the simplistic data generated for use in the previous section, another steppingstone to a full data set is a limited section of real data. To test the efficacy of the proposed algorithms another simple set of data should be illuminating when acquired from a real instrument. To obtain these data an experiment was designed to give results similar to the model data while being actual data acquired by the J105 instrument. A sample was created so as to contain two regions of different sample chemistry. However, to illustrate the use of the methods, the samples used were selected to have some characteristic peaks which would overlap in the spectral



(a) Mass range of two overlapping peaks.

(b) Total ion image of mass range in (a).

Figure 3.6: Zoomed mass rnage of total ion spectrum 3.6a and the corresponding total ion image 3.6b of the silicon oxide and tantalum oxide image (Figure 3.5).

domain, but would be clearly separated in the image domain. Figure 3.5 shows the total ion spectrum and total ion image for a sample comprised of a region of silicon oxide and a region of tantalum oxide.

Looking at a region of the total ion spectrum below 197 Da there are two clear peaks which overlap, Figure 3.5a. Looking at the total ion image for this region gives an image as seen in Figure 3.6b. It should be noted that these peaks arise from two separate spatial locations, one from the silicon oxide and one from the tantalum oxide. The methods described below are applied to all data described above in the results, Section 3.4.

3.3 Methods of compression

3.3.1 Binning

Binning as outlined previously is simply the regular summing of data points, in this case mass channels, to achieve compression. In Figure 3.1 there are three depictions of binning routines applied to a set of peaks based about an arbitrary mass range. The outcomes of these approaches are not ideal for further data analysis or interpretation as information is lost. The loss of peaks distribution information is a major draw back in terms of specific chemical analysis.

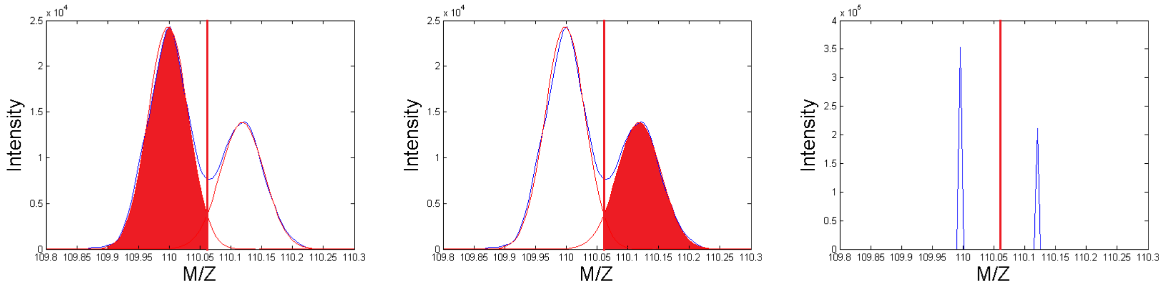
In certain types of MVA each bin/mass channel in the spectrum/histogram are considered as independent of each other, each being their own dimension. When binning routines are applied to single distributions they can be split into two or more bins and thus signal from one distribution are considered as independent variables. Akin to this is when more than one distribution is binned together. These are then considered as from the same distribution and are regarded as such by the multivariate routines. The subsequent approaches described aim to circumvent these problems associated with binning of data.

3.3.2 Selective binning

Here it has been proposed that simple binning of data is not optimal for analysis. A logical step forward would be to group (bin) only individual peaks together. This would reduce data significantly while also not including unwanted noise in the bins from in between the individual peaks as outlined in Section 3.1.3. However, this simple approach does not consider overlapping peaks.

A slightly more sophisticated approach to this problem is to bin single peaks into one bin and with peaks that overlap to use some point between the peaks as a terminating factor. Thus, this reduces the amount of false signal in any one bin but also gives large compression. Consider the examples in Figure 3.7. Here, by choosing a central point, this limits the contribution of peak B in the peak A bin and visa versa. This then reduces the information into two individual bins. This method is comparable to manual selection of limits and is considered as a basis for the following methods.

Peak picking as outlined in the previous chapter is adopted to find peaks. Once peaks are found a peak distribution can be fitted to these features. Using this information a search can be performed to investigate overlapping peaks. All single peaks can be compressed into a single bin. For peaks that overlap the central point between the peaks in question can be used as a limiting factor to determine what is



(a) Selective binning of peak A. (b) Selective binning of peak B. (c) Compressed peaks.

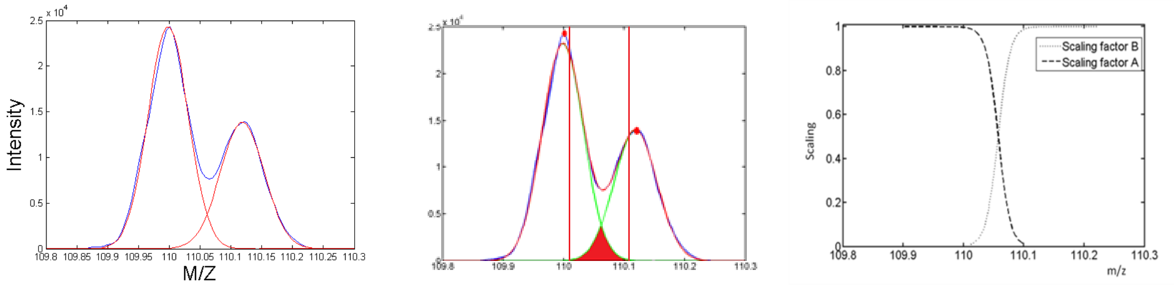
Figure 3.7: Visual representation of the selective binning process.

included in each peak bin. However, instead of just using the central point between the peaks another approach is to use the point at which the two fitted distributions meet. This is chosen as the limiting point because in the case of a large first peak and a small second peak the central point between the peaks is not a realistic limit as the second bin would contain a much greater proportion the first peaks signal than it should. However, the point at which the fitted distributions overlap is the point which will limit the amount of any one peaks signal being included in the other peaks bin. This then will take into account peaks that have large intensity differences. Figure 3.7 depicts an example of this approach. This method can easily be extended to any number of overlapping peaks.

3.3.3 Scaling compression

As can be seen from the previous method it does not deal with the problem of signal that is in the overlap region (the overlap region is represented in Figure 3.8b as the shaded region). The approach simply tries to maximize true signal in a bin and limit the false signal in that bin. Unfortunately, there is not enough information within the total ion spectrum to make a decision about this signal. However, after a fit has been applied to the total ion spectrum, an estimation of the signal can be generated from the fitted peak distributions.

As signal closer to the centre of peak A is more likely to originate from peak A than peak B and the opposite for peak B, a scaling factor which reflects this has



(a) Peaks fitted to the spectrum. (b) Boundaries estimated for peaks. (c) Calculated scaling factors.

Figure 3.8: Visual representation of generating the scaling factor to apply for the scaling compression process.

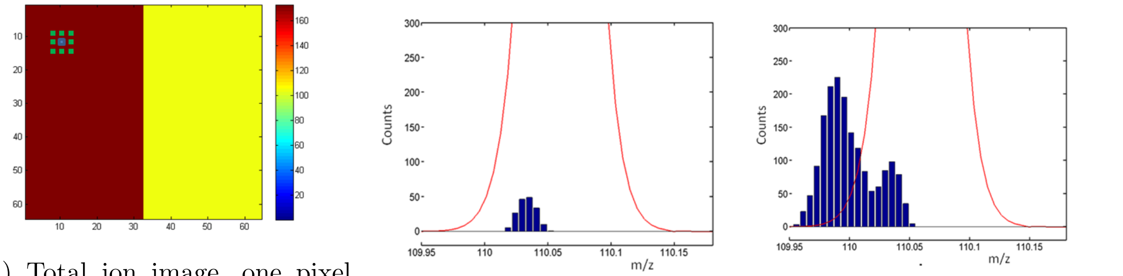
been created to apply to signal in the overlap region in each pixel of the image. Equations 3.1 and 3.2 are created from a fit of a Gaussian to the two overlapping peaks in the data (Figure 3.8a). Since all the signal in the overlap region should be used, these scaling factors are a ratio which sum to one. Each pixel in the image is then scaled using this function to define the proportion of the overlap region that should be attributed to peaks A and B. The scaled signal for A and B can then be summed with the signal outside the overlap region within each individual pixel. This defines a value of peak A and peak B within each pixel.

By scaling all the data across the image an estimation of the actual signal is calculated for each of the peaks. This then can be apportioned to each peak's bin. Clearly this method scales isolated peaks with a factor of 1, so as to use all desired signal.

$$A_{scaling} = \frac{A_{fit}}{A_{fit} + B_{fit}} \tag{3.1}$$

$$B_{scaling} = \frac{B_{fit}}{A_{fit} + B_{fit}} \tag{3.2}$$

Where A_{fit} is the fit of a Gaussian to peak A, B_{fit} the fit of a Gaussian to peak B, $A_{scaling}$ is the calculated scaling factor for peak A and $B_{scaling}$ is the calculated scaling factor for peak B.



(a) Total ion image, one pixel with signal in the overlap region and it's eight neighbour pixels.

(b) Signal in one pixel within the overlap region.

(c) Signal of 3.9b with its eight neighbour pixels.

Figure 3.9: Visual representation of the signal classification process.

3.3.4 Classification of signal

Method two (scaling compression) attempts to deal with the problem of overlapping signal within an image. However, it is still a simplistic approach which includes false signal in both peaks bins. It also reduces real signal from each peak in their particular bin. To combat this a third method has been devised which aims to separate the two overlapping peaks completely. This cannot be done by just incorporating the total ion spectrum, more information is needed. Additional information can be obtained from the spatial information i.e. distribution of signal across the image. By incorporating this information and looking at individual pixels and groups of pixels it is proposed that a decision can be made as to which peak the signal in each pixel came from. Looking across the image pixel by pixel ultimately all the signal from within the overlap region can be separated.

For this new approach consider a single pixel, pixel x, from the model data which only has signal within the overlap region (Figure 3.9a and 3.9b). It is exceedingly difficult to decide from which peak distribution this signal came without prior knowledge. However, by incorporating surrounding pixels, a trend of signal in that local area can be observed. Figure 3.9c is a plot of the signal arising from pixel x in 3.9a and 3.9b with the eight neighbouring pixels' signal also included (Figure 3.9a). Now in Figure 3.9c a clear difference is observed; that there is more signal trending towards peak A than peak B. Using the scaling factors applied in the previous section, a single value for signal in peak A and peak B can be calculated. By looking at the scaled signal under peak A's distribution and the scaled signal under B's distribution the signal from pixel x is assigned to peak A.

Note that a tolerance level of a certain percentage was used here. This tolerance level was set to 20% after empirical testing. A tolerance was used so that if the calculated values for Peak A and Peak B were close, then that signal should not be classified using only this information. In this case, if there remained not enough intensity difference between peak A and peak B in the immediate pixels around pixel x to make a decision, more information is needed to make the decision. In this case the region of interest was expanded again to incorporate more pixels, thus widening the area of the image to aid in the classification. This method is then repeated for each pixel in the image and all signal from the overlap region is assigned to their respective peaks.

However, this method assumes that the nearest neighbours of pixel x will have signal originating from the same distribution, that of peak A for this example. If there is no localisation of signal in the immediate pixels around pixel x expanding the region of interest incorporates signal that is spatially further away from pixel x. This is more likely to bring in signal from the incorrect distribution, i.e. peak B. If this is the case and the signal is incorrectly assigned, that would imply that the true distribution from which the signal in pixel x originated is isolated spatially from other signal arising from that distribution. In real world data there is little information that can be gleaned from such a pixel.

To use this method the assumption that peaks generally localise to certain regions of the image must be made. This will be somewhat dependant on the pixel size. This is an assumption, nevertheless, if these peaks overlap in the total ion spectrum and overlap spatially within the image there is no way to separate them as no other information is available. Thus, since the peaks overlap in the spectral and spatial domains, if the signal is wrongly (impossible to tell) binned to the wrong peak no information has been lost. Therefore, a scaling method may be most appropriate in situations such as this or perhaps the two peaks could be placed in the same bin.

This effect is discussed more in the Discussion, Section 3.5.

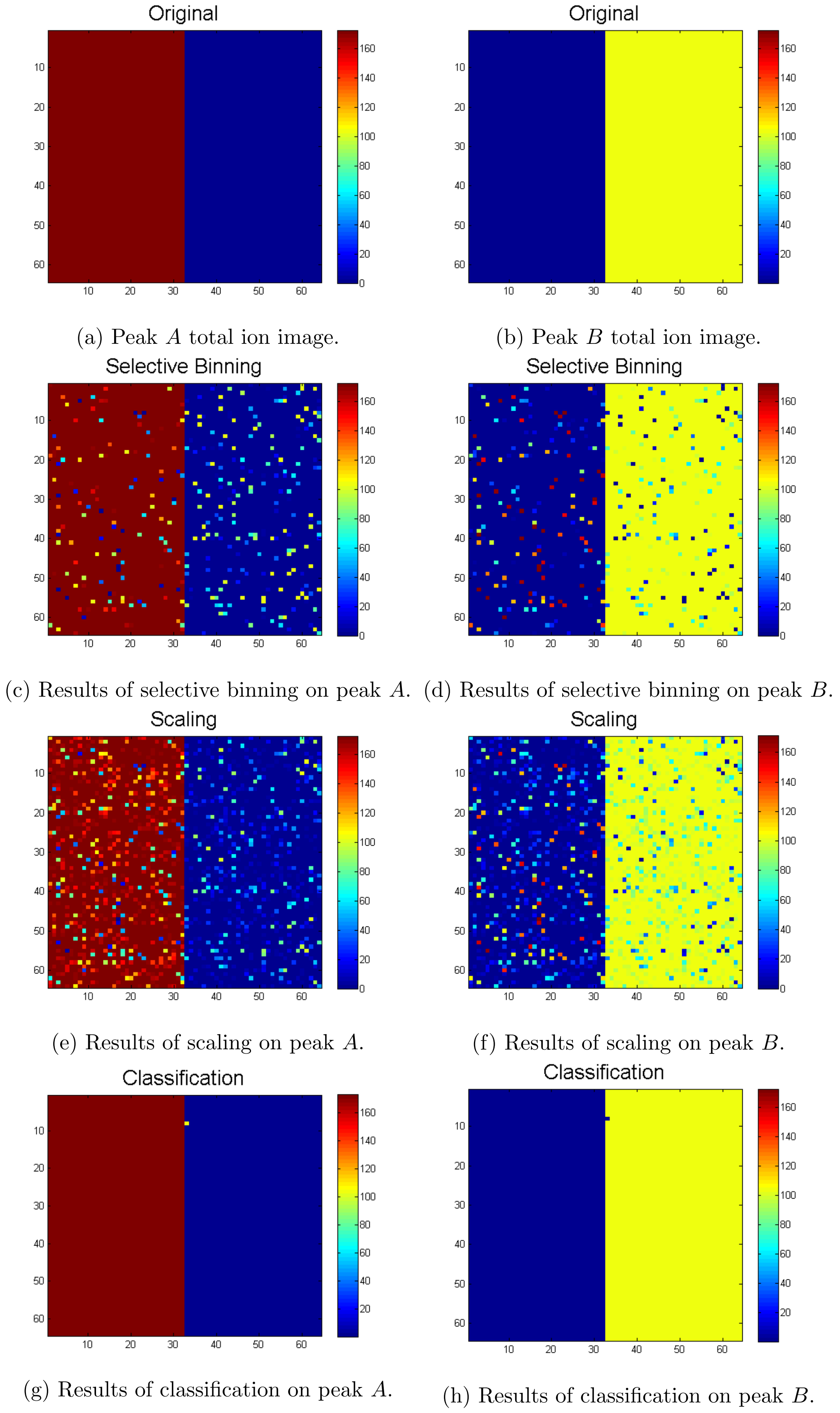
For this method the overlap region is started at three standard deviations from the mean of peak B (to the left of peak B) and ends at three standard deviations from the mean of peak A (to the right of peak A) while the outer regions of both peaks is taken as the point at which the fitted peak reaches a value of 1. This is done so as to ensure that all the information from a given peak is included even if some noise is inserted. This also keeps the overlap region to a slightly more limited range. This is important because if peaks are close together then they will have a lot of overlap, this not only increases the computation time but can also increase the complexity of the problem. These values can be varied depending on user preferences.

3.4 Results

3.4.1 Two Peak Model Data

Figure 3.10 shows the original signal for peak A and peak B and the results of the three outlined methods. Here the desired result it that which most closely resembles the original, Figure 3.10a and 3.10b, as this is the original and true signal. As can be seen selective binning does not give a very close representation of the original data. This is because it chooses a point between the two overlapping peaks to decide where each peak begins and finishes. Therefore true signal is lost to the wrong peak and erroneous signal is gained for that particular peak. This is evident in both peak A and peak B. It is clear that the way the approach was designed that this would be an observed effect.

Selective binning does however perform better than scaling of the data. Intuitively it should be noted that only portions of the data in the overlap region belong to peak A. Similarly for peak B. Thus by scaling all the data with a scaling factor it has a blurring effect, collecting portions of the incorrect peak's signal and "giving away" portions of its true data.



(a) Peak A total ion image.

(b) Peak B total ion image.

(c) Results of selective binning on peak A.

(d) Results of selective binning on peak B.

(e) Results of scaling on peak A.

(f) Results of scaling on peak B.

(g) Results of classification on peak A.

(h) Results of classification on peak B.

Figure 3.10: Images depicting the results of the three approaches for compression of data.

Compression Technique	Peak A percentage	Peak B percentage
Selective Binning	97.34	94.77
Scaling method	94.83	93.82
Classification of signal	100	99.90

Table 3.1: Results of different compression routines on two peak model data.

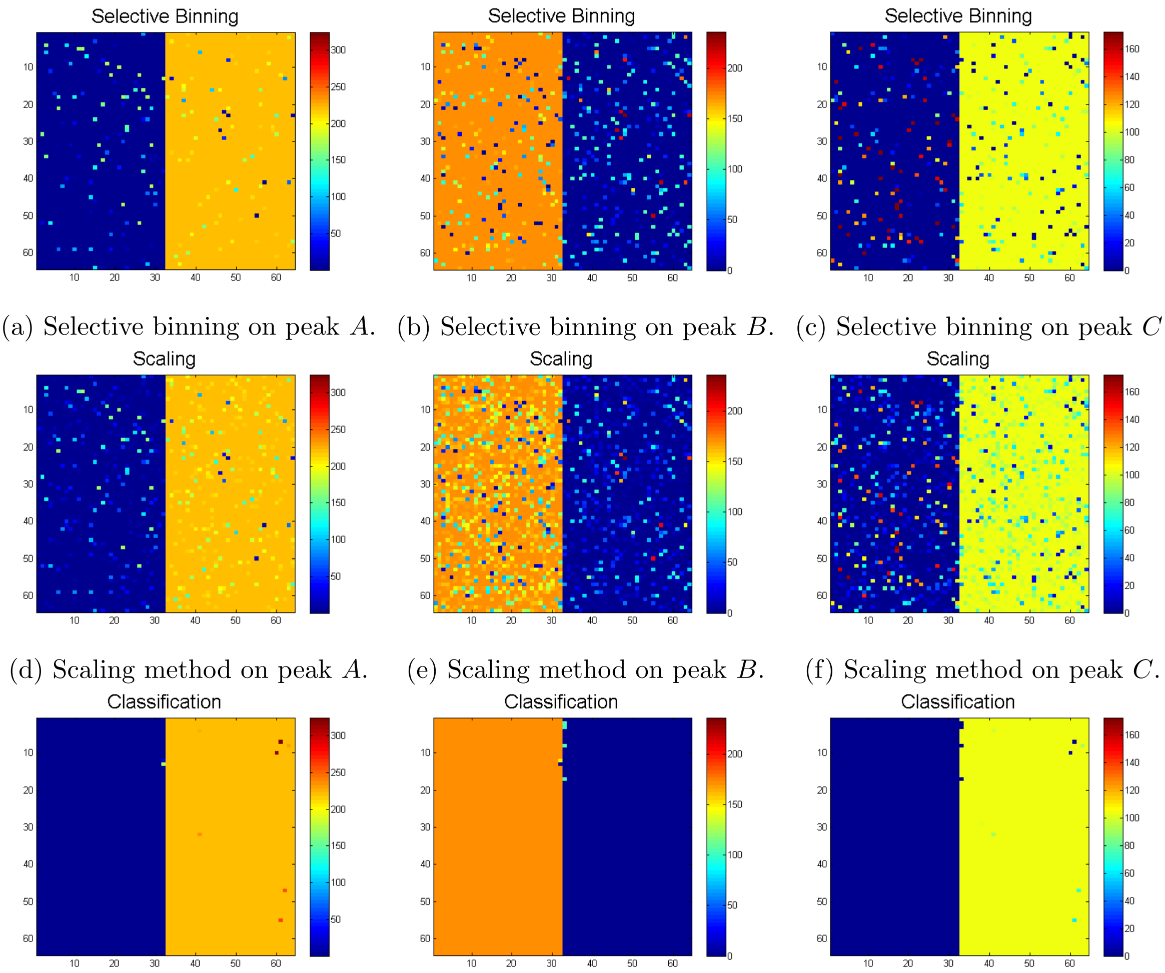
By far the most accurate method for this data is the classification approach. This approach almost entirely re-created the original signal except for one misplaced pixel. Table 3.1 records the percentage of the true signal assigned by each method to the correct peak. Also from this perspective, as well as the visual, the classification method performs with the most accuracy.

From further exploratory analysis reasons can be shown why this pixel was misclassified. Noting the position of this pixel; it is directly on the interface between the two image regions containing the peaks. This has an interesting effect since in every classification of signal in the overlap region, a pixel's signal and its surrounding pixels information is used. This implies that when a pixel is at the interface between two different signal regions the signal from both peak A and peak B distributions are used in the calculation. In the case of the misclassified pixel there was more signal from peak A in the surrounding pixels than there was from peak B. Thus the signal was classified, incorrectly, into peak A.

Another fact to be taken into account is that every pixel containing signal from peak A is more intense than pixels containing signal from peak B. This may have a small effect at the interface between two regions, especially if one peak is much more intense than the other. However, in cases similar to these data this effect is limited to the interface between regions.

3.4.2 Three Peak Model Data

In Figure 3.4 a three peak model data set was been created as described earlier (Section 3.2.1). For these data, not only are there three overlapping peaks but peaks A and C are located in the same pixels spatially and slightly overlap in the



(a) Selective binning on peak A. (b) Selective binning on peak B. (c) Selective binning on peak C.

(d) Scaling method on peak A. (e) Scaling method on peak B. (f) Scaling method on peak C.

(g) Classification on peak A. (h) Classification on peak B. (i) Classification on peak C.

Figure 3.11: Images depicting the results of the three approaches for compression of data.

spectral domain. Figure 3.11 represents the results of the three methods applied to these data for each of the individual peaks.

Accessing the results it is clear that similar outcomes to the two peak model data is observed, with the classification approach achieving the best results and the scaling method achieving the poorest. The scaling method and selective binning routines again fail for the same reasons as noted in the previous section. Table 3.2 represents the percentages correctly assigned for each peak, with the classification approach performing with the highest accuracy.

The classification approach again fails in a few pixels at the interface region. This

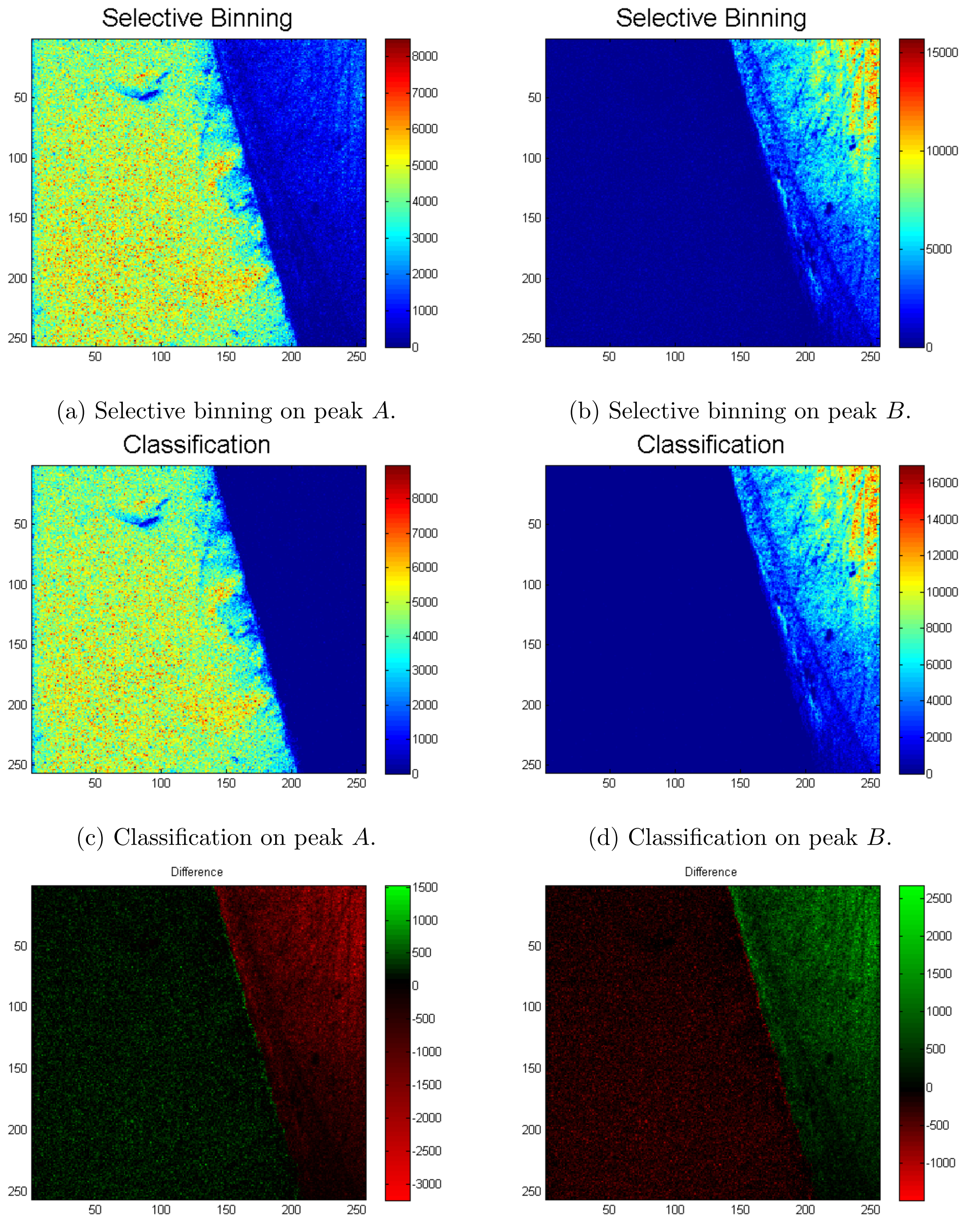
Compression Technique	Peak A percentage	Peak B percentage	Peak C percentage
Selective Binning	99.13	93.77	96.29
Scaling method	98.44	91.30	94.55
Classification of signal	100	99.94	99.70

Table 3.2: Results of different compression routines on three peak model data.

should be noted as a feature of the current design of the algorithm. Unlike the two peak method it can be seen that the classification approach has also failed in classifying some pixels not near the interface. When observing closely between the classification results for peak A and peak C the same pixels have been misclassified. Thus the signal in those pixels have been incorrectly assigned to peak A instead of peak C. It should be noted here that there is some slight overlap between peak A and C in the spectral domain, thus they not only have an overlap region with peak B but also with each other. Therefore there is overlap not only in the spatial domain but some also in the spectral domain. When this is the case the algorithm will classify signal based on the intensity of signal around the pixel in question. In these data peak A is more intense and thus signal that is in this overlap region will be favourably classified into peak A. More on this in the Discussion, Section 3.5.

3.4.3 Real Two Peak Data

The real data acquired from the J105 (discussed in Section 3.2.2) gives another level of complexity over the model data. Figure 3.12 shows the results of the selective binning approach and the classification approach. Also shown in the figure is a difference image for each peak. The difference image is used to highlight the variances between the two approaches. From looking at the results of each method it is hard to distinguish which approach has yielded better results. However, the difference image draws a clear distinction between the two results. The green in the difference images indicate where the classification approach has more signal and the red indicates where the approach has less signal. Again the classification approach has performed much better at separating the overlapping peaks. For peak A there is



(a) Selective binning on peak A.

(b) Selective binning on peak B.

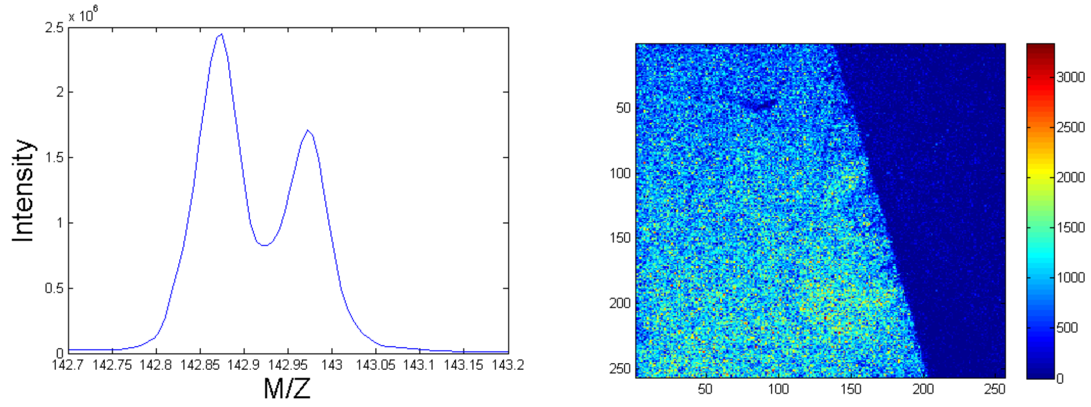
(c) Classification on peak A.

(d) Classification on peak B.

(e) Difference between the methods; peak A.

(f) Difference between the methods; peak B.

Figure 3.12: Images depicting the results of the selective binning approach and classification of signal approach for real data, with difference figures to highlight the discrepancy in signal intensities.



(a) Overlapping peaks near 143 Da in the total ion spectrum.

(b) Total ion image of the region in (a).

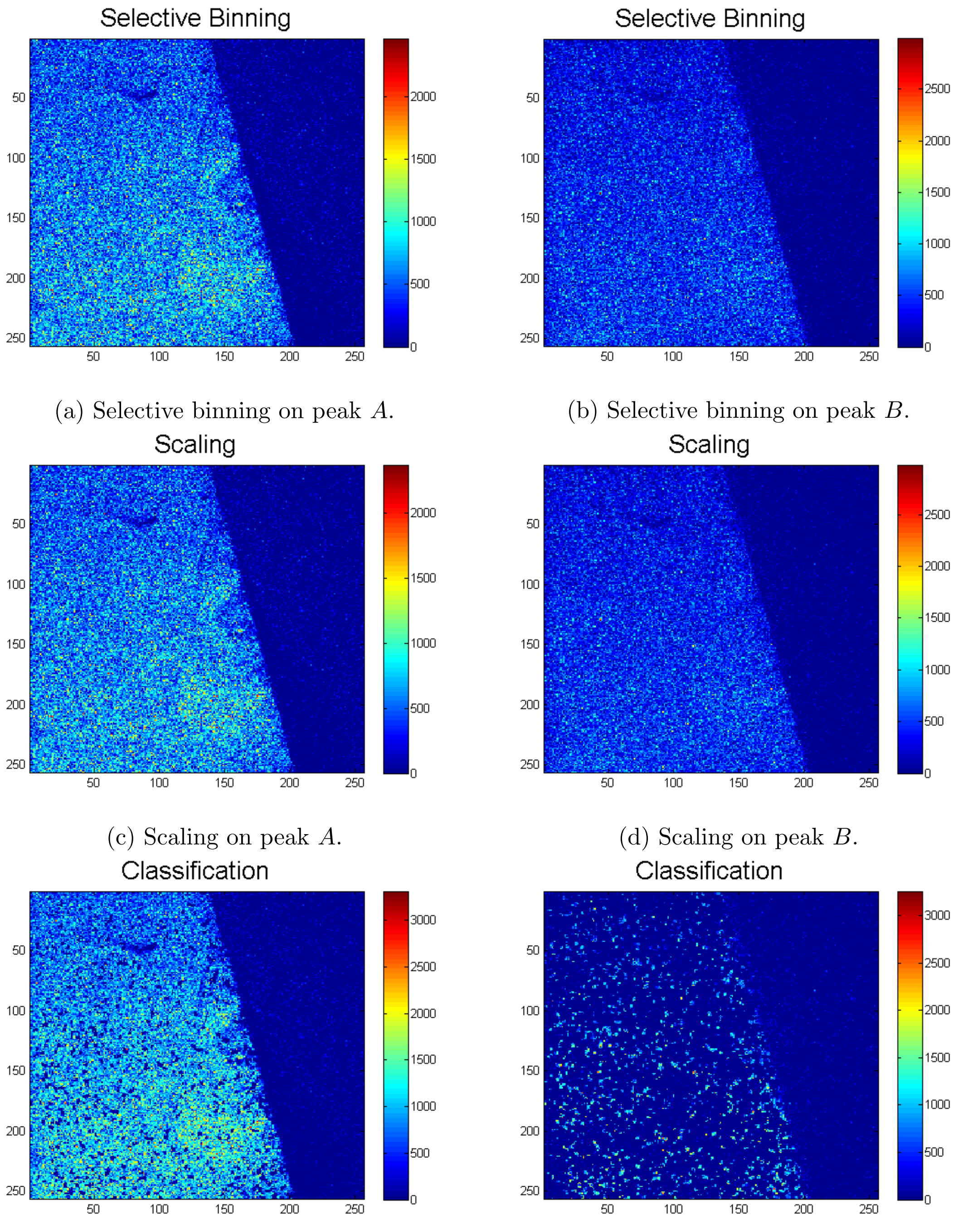
Figure 3.13: Section total ion spectrum containing two overlapping peaks and total ion image for the mass range in (a).

more signal in the left portion of the image the silicon oxide region and less in the right hand region, the tantalum oxide region. Conversely for the second peak. This demonstrates that the classification approach is superior for this style of data where there are localised regions of signal.

3.5 Discussion

As discussed above the classification approach works well when signals are localised spatially. It was alluded to in the three peak model data that when peaks overlap in the spectral and image domains that the classification approach does not perform optimally. Figure 3.13 shows two peaks from the same real data set as used above. Here both peaks overlap in the spectral and also in the image domain.

Figure 3.14 show the results of applying the three methods (selective binning, scaling and classification) to these peaks. From Figures 3.14e and 3.14f it is apparent that the classification method has performed much differently to the other two approaches. As was outlined previously when two peak overlap in the spectrum and the image the signal is preferentially classified to the higher intensity peak by the classification method.



(a) Selective binning on peak A.

(b) Selective binning on peak B.

(c) Scaling on peak A.

(d) Scaling on peak B.

(e) Classification on peak A.

(f) Classification on peak B.

Figure 3.14: Images depicting the results of the selective binning approach, scaling of data and classification of signal approach for real data that overlap both in the spectral and image domains.

For data such as these, the best approach is difficult to ascertain as there is not enough information to decide which peak the data in the overlap region comes from and therefore the true classification cannot be resolved. Here the suggestion is that either the scaling or the selective binning approach be applied to data of this nature. It is clear that both will contain erroneous information, but the idea is to limit this as much as possible. Another approach here would be to bin both these peaks into a single bin for further analysis.

The methods applied here can be directly applied to 3D image stacks. Firstly by summing all the pixels to create an overall total ion spectrum. This then could be peak picked and the methods applied to the images sequentially. Another approach would be to apply the methods to each image independently and then compile the images together again. However, with this approach there may be some complications due to the fact that peaks that are detected may fall into different mass channels across layers depending on the accuracy of both the instrumentation and the peak picking algorithm.

Other alternatives for 3D image stack could be to search not only in the neighbouring pixels in an image but with neighbouring pixels between images. This may give valuable information in terms of localisation as small regions of interest could potentially be probed better by these separation methods with this information. However, instrumental factors may also have an effect on these approaches, such as alignment of the data planes, non-normal incident ion beams, defects in 3D structures etc.

Computation time can be a limiting factor in some cases, in the absence of high throughput analysis for example. Hence, one of the reason that compression such as binning has been classically done. Clearly the time taken to compute each of these algorithms is dependent on the number of peaks in the total ion spectrum but also on the number of overlapping peaks within the spectrum. This is especially true for the classification method as it has to loop over each pixel within the image for each mass channel in all the overlap regions. This as compared to the two other approaches which, for the selective binning is just a simple bin sum once the limits have been set. For the scaling approach it is just simply multiplying all the data by scaling factors and summing the results together. Thus the selective binning and scaling methods are very easily and quickly applied. The time to run the classification method is highly dependant on the image size the number of peaks and the number of overlaps.

3.6 Conclusions

In this chapter some of the challenges and shortcomings of the current norms of compression and imaging in ToF-SIMS image analysis have been discussed. Three new approaches were suggested as alternatives to these approaches. These new methods were evaluated on both model and real data. The proposed method of scaling the data does not give favourable results and thus should not be pursued as an alternative. However, both the selective binning and classification methods gave promising results. The selective binning approach is fast and gives reproducible results superior to current methods. This is suggested as a first step in image analysis.

The classification method gives by far the best results when signal is localised, however in its current implementation it can be slow to compute. This may be a draw back for some but could be a realistic approach were dedicated scripts and machine in place to perform the computation. Moreover, this approach could be used for selected overlapping peaks of interest as opposed to the entire image data. This gives a convenient medium between solely using either approach. The selective binning approach used as default with the classification approach used on selected overlapping peaks.

All the methods described here used automated peak picking as described in Chapter 2 as an input. However, all approaches can be decoupled from the automatic process and could be applied to user defined peaks similarly.

Chapter 4

Classification of SIMS data

Contents

4.1 Introduction

ToF-SIMS has traditionally been seen as a semi-quantitative technique. Though this is true in the sense that the information received from analysis is only a portion of the true total content of the sample, also things such as the matrix effect can affect the results of an analysis, it may be possible to garner quantitative results.

Classical ToF-SIMS machines had to adhere to the static limit to maintain sample integrity. However, with the advent of cluster ion beams it has become apparent that analysis beyond the static limit can be achieved. This has pushed the field into 3D molecular analysis. To capitalize on technique improvements with more quantitative analysis could allow SIMS to become a more mainstream technique for sample analysis.

Though the information gathered in a SIMS experiment is only a portion of the sample, there is still a huge wealth of information received which is characteristic of that sample. Using this information it may be possible to postulate decisive classifications. By using classification algorithms, characteristic traits of particular samples can be learned. These traits can then be used to classify an unknown sample into a given group. This approach has become popular in many MS fields in recent times, with the advances both in the machine learning aspects and the advances in the capability of MS instrumentation [16] [19].

Classically one barrier to this type of analysis have been lack of standardisation between instrumentation, the low number of spectra that it was possible to acquire from a sample and even lower number of samples tested. Due to the limitations of classical instrumentation it was difficult to obtain a useful set of data on which to apply classification methods. However, the large effort in [22] produced data which was fit for purpose. Here these data will again be used for comparison and investigation of classification techniques.

Some classification techniques have been applied to SIMS data previously include Principal Component Discriminant Function Analysis (PC-DFA) [72] and Neural Networks (NN) [62]. To introduce the topic of classification first a brief description of methods to validate classification approaches are outlined.

4.2 Validation

With the development of new learning algorithms and their application into new fields there is a need for accuracy estimation methods. Methods allowing the comparison of several different learning algorithms are important for the performance evaluation. When faced with numerous different classification techniques one needs to test these in terms of their estimated accuracy to make an informed decision about which approach is most relevant to their data or problem.

The general form of validating models/classifiers is to somehow create a 'training set' and a 'test set' (validation set) on which to evaluate the model built by the particular approach. The training set is used to build a model and the model is then applied to the test set to evaluate the models efficacy.

When applying any MVA or classification technique to data scaling is often desirable or even necessary. Scaling approaches are discussed more in Section 4.5. If scaling is applied to data care must be taken to correctly apply to the test set the same scaling as applied to the training data. Note for some approaches all the data cannot be scaled before generating the training and test sets, instead the training set must be scaled first and the test set then must be scaled in the same manner. If this is neglected or misapplied results can suffer greatly.

Here some common approaches for estimating a model accuracy are introduced. For further information refer to [43].

4.2.1 Holdout

Holdout is one of the simplest forms of validation. This is sometimes known as test sample estimate. Here the data are divided into two distinct groups, a training set and a test set. There is no rule enforcing what size each set should be, but general convention is that two thirds of the data are placed in the training set, with the subsequent one third placed in the test set. The training set is used to build a model which can then be applied to the test set. The results of which can be viewed as a percentage accuracy. Training on 66% can reduce the ability to build comprehensive classifiers, especially if the number of samples is small and the number of groups is large, therefore other approaches are often favoured.

4.2.2 K-fold CV

K-fold is a style of validation, which is similar to holdout, where K instances (spectra) are left out of a set for testing. For example, 5-fold cross validation is where the data are randomly partitioned into 5 sets. Of these sets, one is retained as a test set. The cross-validation process can then be repeated another 4 ($K - 1$) times. Thus, each of the 5 sub-samples is used only once as the validation set. The 5 results give a coarse estimate of the mean can then be averaged to return a single estimation of accuracy. The advantage of this approach is the fact that each sample is used once in the validation. Thus, each instance (spectrum) from the original data is predicted once. Meaning, that the cross-validation accuracy is the percentage of the original data which was correctly classified.

4.2.3 Leave-One-Out-Cross-Validation

Leave-One-Out-Cross-Validation (LOOCV) is a method whereby all the data except for one spectrum is kept as the training data and the test set comprises of a single spectrum. This process can then be repeated for each spectrum. This can be considered a special case of K-fold where K is equal to the number of spectra in the original data. This results in each spectrum being used only once in the test set and the results can be seen as a percentage of the original data which were correctly classified.

For each iteration of LOOCV almost all the data are used in the training set, this has the advantage for small data as it allows models to be built that are close to the models of the full data. However, for large data this approach may be time consuming and thus not an attractive approach.

4.2.4 Bootstrapping

Bootstrapping, introduced by Efron (further detailed in [18]), is a method employed to assign measures of accuracy to classification estimates. This is achieved by sampling a subset of the data the same size as the original data **with replacement**. Thus, if there is n spectra in the original data, the Bootstrap set will also have n samples. Given the fact that bootstrapping is done with replacement the probability that any instance is chosen in the training set is 0.632. Therefore the probability that an instance is in the test set is 0.368.

To generalise the results of Bootstrapping it can be performed repeatedly, thus building up statistics. From these repeated test, statistics such as the mean and standard deviation of the classifier can be obtained.

4.2.5 Stratification

Stratification is a process whereby each class is considered independently by the generation of the training and test sets under a given approach. For example in K-fold, instead of dividing the total data into K-folds, stratified K-fold would create K-folds from each class and then concatenate them to create the training and test sets. This approach of stratification can give less biased results [43], especially if the differences in class sizes is large or one or more classes is a small set of samples.

4.3 Classification Techniques

4.3.1 PC-DFA

PCA (Section 1.3.6) is used here as a dimensionality reduction technique to input into Discriminant Function Analysis (DFA). DFA is a method applied to decipher among sets of samples from differing distributions or populations. Generally in the SIMS community DFA has been the conventional terminology used but DFA is also addressed as Canonical Variates Analysis (CVA), here the convention is continued and DFA will be used.

PCA has become a staple of chemometric analysis. PCA is a dimensionality reduction tool. As such it can play a vital role in a diverse set of problems. In PC-DFA PCA is used as a dimensionality reduction tool which is then modified in the DFA step to create a sub-space where classes of samples are separated in this sub-space. PC-DFA has been applied numerous times in the SIMS community with good results [2][70].

PCA can be considered as an unsupervised method for identifying patterns in data. The goal of PCA is to identify trends or patterns, through variance, in data and expressing the data in terms of the maximum variance. DFA, a supervised technique then calculates a linear composition of the PCA variables to create a weighted output of these. The goal of this analysis is to maximise between group variance while minimizing within group variance. Thus, in the space in which the instances are projected they are well separated in space according to class. This is achieved by maximising the Fisher ratio. By applying DFA in effect the PCA results are warped into a new space where groups are more easily separable. By doing this, in contrast to PCA, the DFA results are not orthogonal. From here classification can be applied.

Here a distance metric is used to calculate to which distribution the test data are closest. The Mahalanobis distance metric is used here. Using the Mahalanobis distance is preferable to say city block/ Euclidean distance as it takes into account the multivariate distribution of the data and is scale invariant. The test data are classified to the group/class to which it is closest in terms of the Mahalanobis distance. The Mahalanobis distance is given by:

$$d_x = \sqrt{(x - \overline{G})\sigma_G^{-1}(x - \overline{G})'} \tag{4.1}$$

Where d_x is the Mahalanobis distance of spectrum x,$\overline{G}$ is the mean of the group of target points and σ_G^{-1} is the inverse of the covariance of group G.

With PC-DFA there are two tunable parameters; the number of principal components to use as an input to DFA and the number of discriminant functions to calculate. Here however the number of discriminant functions is limited either the number of groups minus one or the number of dimensions. In general 10 principal components retains a very high percentage of the variance within the data. In [22] three PCs were used as an input to DFA to ensure the model did not overfit the data. Some methods have been proposed for selection of relevant number of PCs to used such as a PRESS test or a scree plot. In Section 4.6 the number of input PCs is kept constant to track efficacy of different pre-processing techniques and to allow for direct comparison of results.

4.3.2 AdaBoost

AdaBoost, introduced by Freund and Schapire [27] is an ensemble method for constructing a “strong” classier as combination of weak learners or classifiers. By choosing a defined set of weak classifiers, these then sum together to make a final decision classifier. AdaBoost was chosen as a method of analysis for SIMS as it relies on simple weak classifiers to build a strong classifier. a weak classifier can be intuited as a rough rule of thumb, or a rule that is right most of the time. This is attractive for SIMS due to the variability in SIMS data, even on samples with the same chemistry. Thus, as AdaBoost learns with simple weak classifier, it reduces the need for having a decisive difference between two classes, instead the subtle trends can be used to learn the differences. Below a brief description of the AdaBoost algorithm is given, for a more complete description refer to [28].

Suppose hypothetical data as in Figure 4.1. Where the ’+’ is representative of one class and the ’-’ are representative of another. These two dimensions can be considered as two individual mass channels in a SIMS spectrum, with the value along each axis the intensity of that signal. The next step is to learn these samples so that the model can classify new unseen hypothetical spectra.

In the case of AdaBoost it looks to find a simple classifier that will correctly

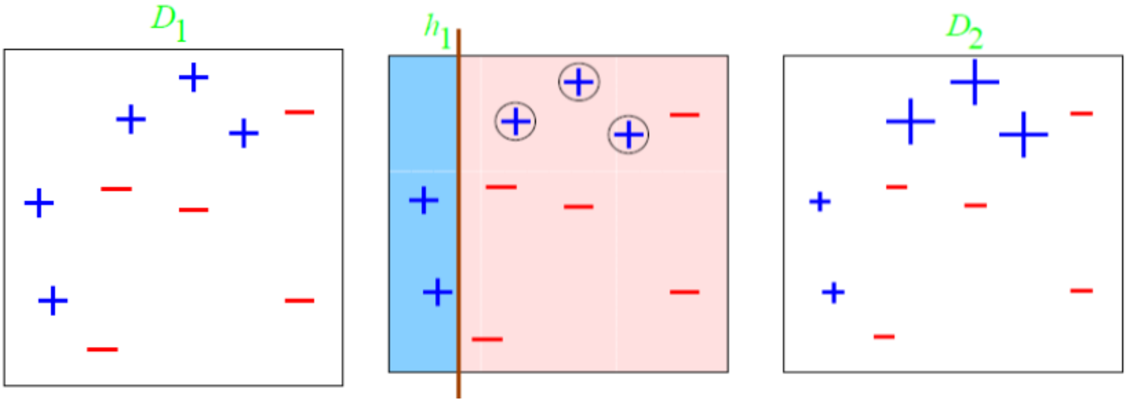


Figure 4.1: Graphical representation of the Boosting process, depicting some model data. Reproduced from [64].

classify as many of the samples in one go as possible. Each sample has the same weight to begin, meaning that each sample is as important as every other sample. This is important to note as AdaBoost seeks to minimize the weighted error of the samples. In the case of this hypothetical data the weak classifier is calculated to be the line h_1 in Figure 4.1b. As can be seen from the figure, all the samples to the left are now classified by this weak classifier as positive samples and all samples to the right are classified as negative. This gives an error of 30% as three of the '+' class are misclassified.

AdaBoost now takes this results and re-weights all the samples accordingly. Correctly classified samples are down weighted and incorrectly classified samples are increased in weight, Figure 4.1c. Once this is done the algorithm again tries to classify using a weak classifier. This time however, the weighting of each sample is not the same. Here AdaBoost calculates the weak classifier that minimises the weighted error. Therefore, it is no longer looking to just classify samples correctly, it is preferentially trying to classify the samples that were previously misclassified.

AdaBoost calculates the second weak classifier as h_2 in Figure 4.2b. Here it has now classified the rest of the '+' class correctly. However, it has now misclassified three '-' class samples. These samples are now re-weighted again, Figure 4.2c. On the third round of boosting AdaBoost calculates a third and final classifier that reduces the error to zero Figure 4.2d.

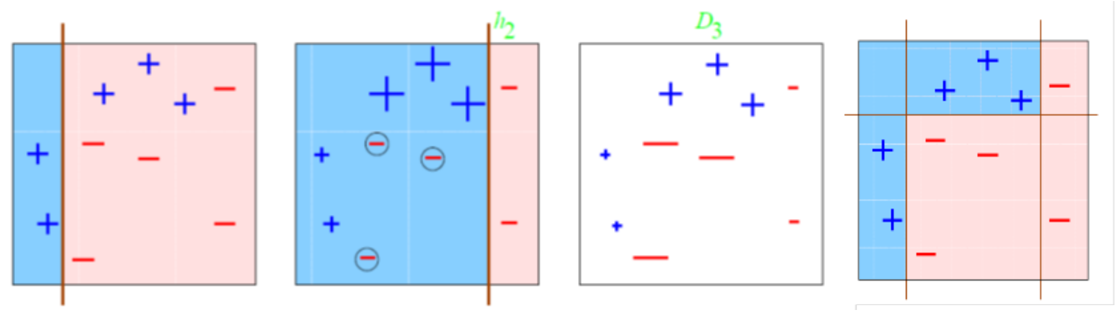


Figure 4.2: Graphical representaion of the Boosting process. Reproduced from [64].

This process of applying weak classifiers and re-weighting the samples is continued as many times as the user desires, or generally when the error reaches zero. For unseen data, each weak classifier is applied to the data, a final score is then assigned to that unseen spectrum for example. This final score is a linear (weighted) combination of all the weak classifiers. The higher the final score the more likely it is to be correctly classified. In a two class case, higher positive score means the sample is classified in the positive class, conversely for the negative scores. The closer to zero the score the more uncertain the output. One of the clear advantages of AdaBoost other than simplicity is the fact that parameter tuning is not a problem. Compared to SVMs (discussed in the next section), where parameter tuning is highly important and is not trivial to select the best tuning parameters for a given kernel. Also tuning of these parameters can be time consuming, AdaBoost avoids these.

AdaBoost classification returns a value for a particular classification. By looking at the values across all the classes an estimate of the models certainty on a particular classification can be estimated. Also with AdaBoost an insight into the classification approach can be gained by looking at the weak classifiers. Thought the weighting of each spectrum will change at each round of boosting, all spectra are equally weighted for the first round of boosting. Thus, by looking at the top N results for the weak learner the mass channels that are instructive for classification can be observed. This is discussed more in Section 4.7.

4.3.3 SVMs

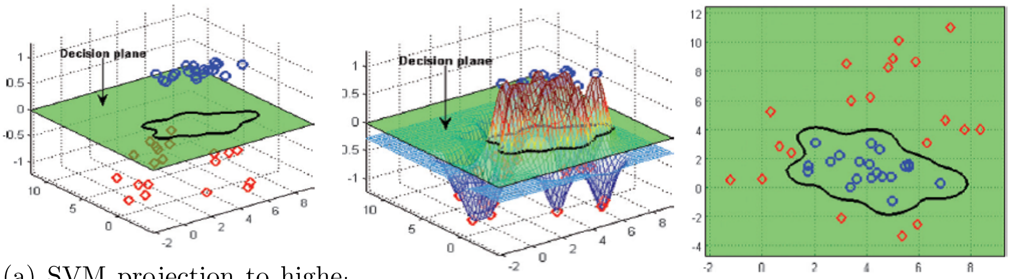
Support vector machines (SVMs) are a group of supervised learning methods that can be applied to classication or regression. SVMs were originally proposed by Cortes and Vapnik [14] and have become a standard of classification within the Machine Learning community. SVM applications have been applied to numerous areas including bioinformatics [30] and mass spectrometry [82]. SVMs were originally defined for the classication of linearly separable classes of objects but have been extended for non-linear applications. Below a brief description of linear and non-linear (Kernel) SVMs are introduced. The routines applied here are the LIBLINEAR [20] and LIBSVM [11] implementations which are freely available and contain MATLAB interfaces through Mex files. For a more in depth description of SVMs and their applications in bioinformatics see [8].

Linear SVMs

Linear SVMs aim to find the unique hyperplane, or line if data are two dimensional, having the maximum margin between two classes. Written in its simplest form the linear boundary is given by:

$$g(x_i) = sgn(wx_i' + b) = sgn(b + \sum_{j=1}^{j} w_j x_{ij}) \tag{4.2}$$

where w and b are called weight and bias parameters that are determined from the training set, x is an individual sample. The sign of g determines which class an unknown is classified to, positive classifies as one class and negative the other. This classification function finds the hyperplane (w, b) that can be defined by coordinates x satisfying the condition $wx' + b = 0$ which divides the dataspace into two regions opposite in sign.



(a) SVM projection to highe dimensions.

(b) SVM decision plane.

(c) SVM non linear classifier.

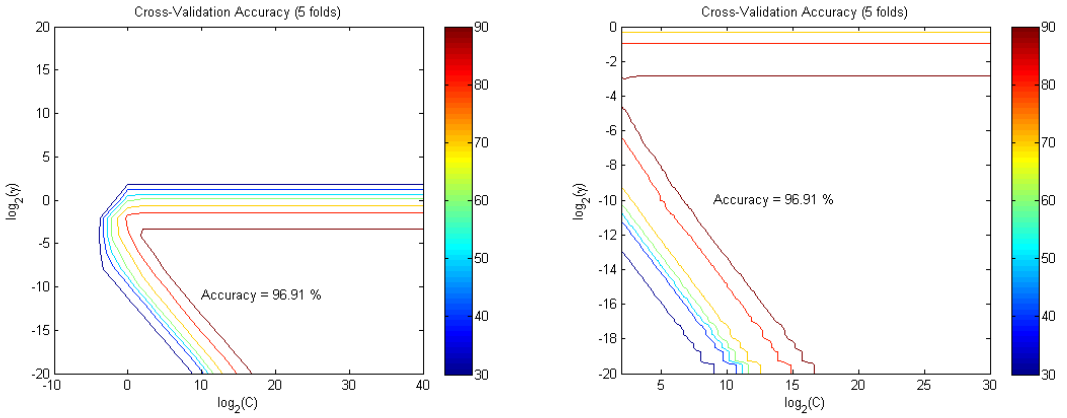
Figure 4.3: Graphical representation of the SVM process. Reproduced from [8]

Kernel SVMs

SVM can also be used to separate classes that cannot be separated with a linear classier, thus non-linear. In these cases, the coordinates of the objects are mapped into a feature space using non-linear functions. The feature space is a higher dimensional space than the space of the input samples. This higher dimensional space then allows the two classes to be separated with a linear classier (Figure 4.3). This typically produces a non-linear classification boundary in the original input space. The dimension of the transformed space can be very large, but through sophisticated mathematical approaches this undesirable effect can be negated.

The cookbook approaches outlined by [40] suggest a straight forward approach for analysis. The steps postulated for analysis are:

- Transform the data to a sympathetic format
- Conduct a simple scaling.
- Conduct a simple scaling.
- Use cross-validation to find the best C and γ parmeter.
- Use the best C and γ to train across the whole training set.
- Apply/test the model....



(a) Wide grid search

(b) Narrow grid search.

Figure 4.4: Grid searches.

Cross validation grid search

There are two parameters that need to be tuned for an RBF kernel; C and γ. It is not known which C or γ is best for a given problem. Thus, a parameter search is implemented here. The aim of this is to ascertain the most applicable parameters to train the proposed model so that the classifier model can accurately predict unseen data. A common utilised approach is to perform N-fold cross validation with a range of parameters. This will create a two dimensional parameter space. The grid search thus trains and test the data using N-fold cross validation for each point in the parameter space. The prediction accuracy obtained from the test set with the highest accuracy reflects the performance on classifying an unseen data set. A grid search can be time consuming, therefore using a coarse grid first to find an applicable range a useful approach. After identifying a "better" region on the grid, a finer grid search on that region can be conducted . Figure 4.4 depicts a coarse and fine grid search for C and γ. After the best C and γ parameters are found, the whole training set is trained again to generate a final classifier.

For cases such as the data here, i.e. small sample size with large feature size the SVM, has been noted as a good tool. In [40] they note that on data such as these that the linear form of the SVM can perform as well as the non-linear as the data are already high dimensional thus a projection to a higher space may not be needed. Therefore, non-linear mapping does not improve the performance. The linear SVM has the advantage that only the C parameter needs to be searched. It also should be noted that the linear SVM is a special case of the RBF kernel, therefore the RBF should perform at least as well as the linear once the parameter space has been searched appropriately. Here both methods are test on the data to compare performance on SIMS data.

4.4 Data

The data used for this investigation are UTI bacterial data previously studied by PC-DFA [22][72]. The data were acquired using the BioToF instrument [7]. The data consist of 6 different bacteria: *E. coli, Klebsiella oxytoca, Klebsiella pneumonia, C. freundii, Enterococcus* spp., *P. mirabilis.* For further details on the data refer to [22]. Each bacterium will be considered to be one class. Thus, there are six classes in this classification investigation. A total of 163 spectra were used for the classification study across the six classes.

Comparisons are made in terms of scaling, binning and mass ranges of data using some of the validation methods outlined previously in the results, Section 4.6.

4.5 Scaling

Scaling of data is an important step in any analysis of data. The outcomes of a given approach can be changed dramatically due to the use of scaling. The main objective of scaling is to avoid features that have greater intensity ranges dominating those in smaller numeric ranges. However, by applying some scaling it can have an adverse effect. This is especially true within MS type data where there is a large proportion of the spectrum which is noise. Thus scaling or any pre-processing must be carefully considered.

The same pre-treatment must be applied to all spectra in the training and test sets. Thus if the training data are scaled to max counts, the same scaling factor must be applied to the test data. Data pre-processing can consists of several different approaches from peak selection, probabilistic weighting, peak intensity scaling, and weighting of peak intensity due to its mass position. Below different types of scalings are discussed and their application to classification problems. Poisson scaling was performed on all data prior to scalings outlined below.

Binning

Due to the variability in SIMS data they are often binned for computational simplicity and size reduction. Here data are binned across several different scales to test the effect it has on classification routines. Data are binned to 0.1, 0.25, 0.5 and 1 Da. Here the convention of binning around the nominal mass was observed.

Square root scaling

Spectra obtained from time of flight mass spectrometers tend to have large peaks in the lower mass range, this is due to the nature in which the spectra are recorded. Since higher mass values are often more descriptive it can be instructive to pre-treat the spectra by taking the square root of the intensity values (Equation 5). This is done as higher mass values tend to have much less intensity. This has the effect of both reducing the relative power of the larger low mass peaks, thus increasing the power of the less intense higher mass peaks. Also the cube root and higher powers can be applied.

Mass intensity scaling

Another form of pre-treatment is multiplying each intensity by its corresponding m/z value raised to a power (Equation 6). This again greatly increases the significance of values with a high mass and thus decreases the significance of lower mass peaks.

Normalise max counts

To normalise to max counts each mass channel in the spectrum is divided by the maximum number of counts for that mass channel, thus each mass channel becomes a value between 0 and 1.

Autoscaling

Mean centre and divide by standard deviation of each mass channel. The result is that each column of has a mean of zero and a standard deviation of one.

Vector normalisation

Vector normalisation is the normalising of each element of the vector by the length of the vector. This gives vector which have unit length.

Normalise total counts

Each mass channel is divided by the total number of counts, thus the entire spectrum sums to 1.

Results comparing these scaling techniques can be found in the results, Section 4.6.

4.6 Results

4.6.1 Linear vs. Kernel SVM

To compare a linear SVM and a kernel SVM here we apply both methods to the data outlined earlier. Using leave-one-out cross validation to test the efficacy of each approach. By applying leave-one-out a direct comparison on the misclassified spectra can be observed.

To test each approach the data are scaled to a 0-1 range in keeping with procedures outlined in Section 4.3.3. Thus each mass channel in the training set is scaled by the max value in that mass channel across all the training spectra. The test set is then scaled using the same values to scale as the training set. For each iteration a grid search across the C and γ parameters for the kernel SVM, the C value for the

Applied Scaling	Classification technique		
	PC-DFA	AdaBoost	SVM
No Sclaing	77.30	85.28	26.38
Scaled $0-1$	77.30	85.28	95.71
Square Root	77.30	85.28	-
Square root and scaled $0-1$	76.07	85.28	96.32
Cube root	76.69	85.28	-
Cube root and scaled $0-1$	76.69	85.28	96.93
Data*mass and scaled $0-1$	77.30	85.28	93.32
autoscale	77.91	85.28	93.25
vecnorm	79.14	88.34	95.71
Max counts	79.75	88.34	95.09
Total counts	78.53	91.41	95.71

Table 4.1: Results of LOOCV for various scalings.

linear SVM, is performed to ensure maximum accuracy. Input data were binned to 1 Da as detailed above for ease of analysis.

For leave-one-out cross validation the linear SVM achieved 92.02% correctly classified, while the RBF attained 95.71%. 5-fold cross validation results followed the same trend. It was noted in [40], for data such as the data used here where the number of samples is much less than the number of features, the linear classifier performs well and it may not be necessary to map to a higher dimension but that the RBF is at least as good as the linear after the parameter space has been searched. In this case the RBF outperforms the linear classifier. For the subsequent experiments the RBF kernel is used.

4.6.2 Scaling

Scaling of the data can be an important pre-processing step in classification. Here an investigation into the efficacy of several different scaling approaches was carried out. Each scaling was applied both to the training and test data as outlined in Section 4.5. A leave-one-out cross validation approach was adopted to investigate these approaches. This was done so that direct comparison between the effects of the scaling routines could easily be interpreted.

Table 4.1 shows a list of the different scalings applied to the data with the corresponding percentages for the LOOCV. These experiments were all run on the same data i.e. as described above and binned to 1 Da. To find the parameters for the RBF kernel SVM a grid search was again applied at each iteration. For AdaBoost and PC-DFA the parameters were kept constant. With 10 principal components and 5 discriminant functions used for PC-DFA so as not to overfit the data and the number of weak classifiers was set to 20 for AdaBoost.

Having tested different scaling possibilities, it is clear that no single scaling routine generates the best results for the different classification approaches for this data. Though the scaling acquired diverse results, with the exception of no scaling for SVMs, the scalings had little effect on the overall percentage results. However trivial the increases may seem, on larger data these effects could be compounded.

From the results it should be noted that for SVMs the data needs to be scaled to a small range such as -1 to 1, or as was generally the case here from 0 to 1 as noted in Section 4.3.3. This scaling has a dramatic effect on the classification ability of the SVMs as can be noted from first results for SVM in Table 4.1. This is the reason that two results are absent for the SVM as the goal was to test the effect of the 0-1 scaling on AdaBoost and PC-DFA. Scaling 0-1 it has no effect on AdaBoost, either in classification percentage or which spectra were misclassified. From the table of results the square root and square root scaling which is again scaled 0 to 1, this has an effect on the PC-DFA. Though the classification results themselves do not vary much the actual spectra which were incorrectly classified do change somewhat. This effect was also observed with the cube root scaling with and without scaling the data from 0 to 1.

For PC-DFA scaling the data to max counts produced the best results, while for AdaBoost scaling the data to total counts produced the highest percentage, with cube root scaling for SVM giving a slight percentage increase.

Applied Scaling	Classification technique		
	PC-DFA	AdaBoost	SVM
Binned 0.1	55.21	75.46	89.96
Binned 0.25	59.51	80.98	91.41
Binned 0.5	63.19	88.96	92.64
Binned 1	78.53	91.41	95.71

Table 4.2: Results of LOOCV for various binning scales.

4.6.3 Binning

Here the classification routines were tested with SIMS data binned to different mass ranges. Binning is a common approach adopted throughout the SIMS community. The effect of binning can have both positive and negative outcomes depending on ones perspective. Intuitively it would seem that classifying on full resolution data would be more instructive as the true signal acquired can be classified and individual peaks can be resolved to do this. However with this said full resolution data also contain a majority of mass channels that contain noise. These mass channels are considered by the classification routines and could be used to build the classification model. The advantage of binning here is that not only is true signal summed together, but the noise is summed. Binning over different scales could average the noise out and thus reduce the likelihood of noise interfering with the classification. Table 4.2 shows the results for LOOCV on the data with scaling to total counts kept as the scaling routine across all binning approaches.

Observing Table 4.2 there is a direct correlation between binning of the data and the ability of the classification routines to correctly classify the samples. Here the smaller the bin size the lower the accuracy of each classification routine.

4.6.4 Mass Range

In classical SIMS the majority of the signal acquired appears in the lower mass region of the spectrum. Thus the higher up the mass scale the smaller the signal to noise ratio. At the lower end of the spectrum salt peaks can dominate. The case was made

Applied Scaling	Classification technique		
	PC-DFA	AdaBoost	SVM
1-1000	78.53	91.41	95.71
50-1000	82.82	90.18	94.48
100-1000	78.53	87.73	93.25
50-300	84.05	91.41	95.09
50-400	83.44	92.64	95.71
50-500	76.69	90.18	95.71
50-600	77.30	90.80	94.48
50-750	77.91	90.18	94.48

Table 4.3: Results of LOOCV for various mass ranges. All binned to 1 Da.

for removing the first 50 Da in [22] to exclude these salt peaks which can dominate spectra and are less molecule specific. For these reasons here an investigation into the effect of pre-selecting mass ranges for classification is carried out.

From the Results in Table 4.3 the selected mass ranges over which the spectra were tested resulted in differing percentages with no decisive pattern. There is somewhat of a pattern to be observed, all three algorithms achieve their highest results over smaller mass ranges, $50 - 300$ and $50 - 400$. No one mass range clearly dominates in terms of percentage correctly classified. However, the 50-400 range did achieve the highest results for both AdaBoost and SVM.

It is interesting to note the results of PC-DFA for $1-1000$, $50-1000$ and $100-1000$. Here the classification jumps once the first 50 Da are removed but drops again once the next 50 Da are removed. Presumably the classification is being forced to classify on a smaller range when mass ranges are removed and thus the improvement in classification could be understood, but also possible discriminatory information is also being removed thus the drop in classification. Also the other routines do not suffer in such a dramatic way but their percentages do drop over these mass ranges also.

Classification Routine	Mean percentage	Standard Deviation
PC-DFA	77.48	4.66
AdaBoost	84.21	5.32
SVM	92.69	3.54

Table 4.4: Results of 500 Boostraps for classification routines.

4.6.5 Bootstrapping of models

To test the validity of generated models. Bootstrapping with substitution as described in Section 4.2.4 was applied to the data. 500 Bootstraps were run for each set. Stratification was used here to ensure that each class had an appropriate number of spectra to train and test. Table 4.4 details the results of these experiments. For these experiment the data were scaled independently for each classification routine. With the highest scoring scaling being applied to each approach as estimated by LOOCV in Table 4.1.

From Table 4.4 SVMs achieve the highest performance. They attain the highest classification percentages but also the smallest standard deviation in comparison to the other two methods.

4.7 Discussion

The experiments carried out here verify that it is possible to attain very high accuracy classification results with SIMS data. However, extracting chemical information from the classifiers to decipher how a certain classifier model computes a results is not straight forward. Since the SVMs used here project to a higher dimensional space little information can be extracted. Distance scores can be returned describing the distance of the unknowns to each of the support vectors. In the case of PC-DFA the PCA scores and loadings can be interpreted and the DFA loadings can also be looked at for peaks of interest.

AdaBoost is based on weak classifiers which can easily be visualised. However, at each round of Boosting the weightings on each spectrum change and the algorithm

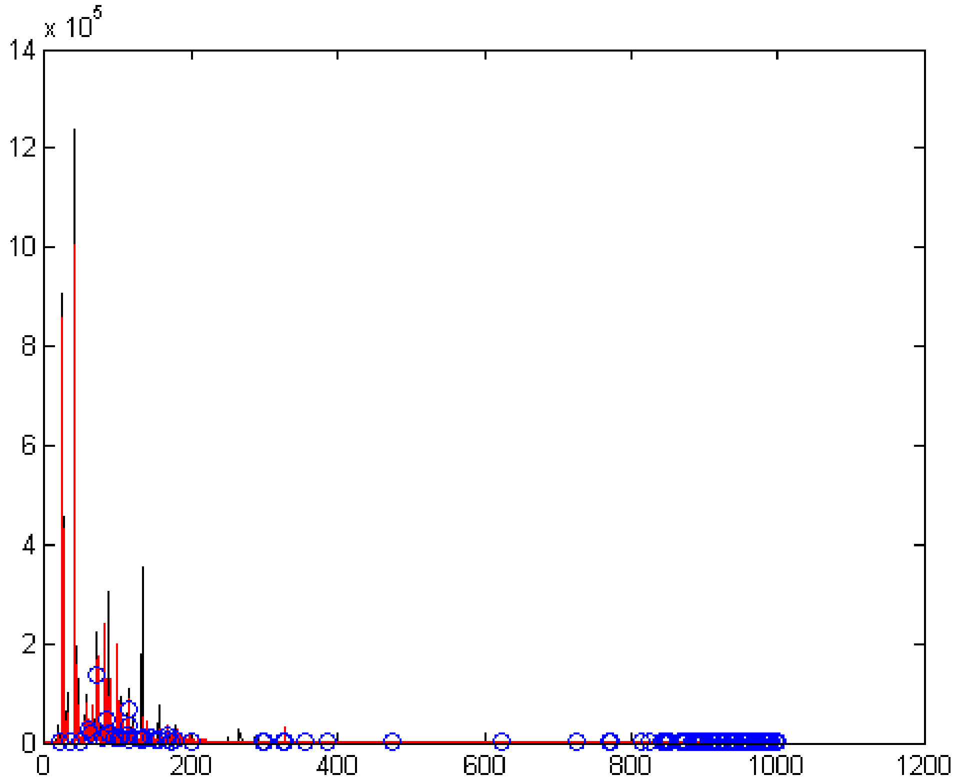


Figure 4.5: 100 decision stumps (weak classifiers) for two classes (Red and black). The blue circles indicate the weak classifiers.

tries to preferentially classify. Therefore the weak classifiers after the first round of boosting become very difficult to interpret. However, by looking at the top N weak classifier in the first round of Boosting when all the spectra are weighted equally some insight into the differences between classes and how the algorithm works can be gained. Figure 4.5 shows the top 100 weak classifiers for a two class set. The red spectra are one class and the black spectra another class. The blue circles represent the weak classifiers. From this figure it can be seen that there are a lot of weak classifiers at very high mass range thus these are mass channels which are "good" for classifying or at least seen as better than the rest by AdaBoost. There are also lots of classifiers in the lower mass ranges. This is a Univariate style of analysis and could be instructive to an analyst in discovering differences in chemistry.

One of the limiting factors of this study was the small class sizes as an input. To achieve higher classification results more data may be needed. Though the sample sizes were small the proof-of-concept has been realised, that SIMS data can give decisive classification results. To expand on this type of analysis requires a large effort in terms of experimentation. However, with the recent advances in duty cycle of machines like the J105 larger data can be investigated.

Classification of image data is another degree of complexity. The information contained in any single pixel may be too little to train an accurate classifier. Conversely the total ion spectrum may contain information from too diverse a set of pixels. Problems due to sample size and sample stubs can play a large role here. The number of pixels analysed that contain the particular sample to be tested. Also different ion doses will produce different results, with the transmission of the machine or tuning of the machine playing a large role also. With some of these effects it may be possible to scale out the effects, but this is still only speculation. There are still many factors which need to be addressed before large scale classification routines can be applied to SIMS images. However, with the proof of concept achieved here there is clearly scope for this type of analysis to progress.

4.8 Conclusion

Here classification techniques were applied to ToF-SIMS data. Results of classification clearly show that SVMs perform the best on these data. However, reasonable levels of accuracy were achieved by both PC-DFA and AdaBoost. Validation techniques confirm that SVMs perform with the highest accuracy. Scaling routines investigated suggested that there may not be one optimal approach for these data, but that each classification routine perform optimally on different scalings.

Chapter 5

Data processing and interpretation toolbox

Contents

5.1 Introduction

As introduced in Chapter 1 ToF-SIMS has become a sophisticated analytical tool. Evaluating, storing and processing the subsequent data is increasingly a challenge due to the nature of the data acquired. One of the most fundamental tasks in this line of investigation is to understand the resultant spectra observed from the analysis. However, with these large datasets and indeed the complexity of some spectra observed, increasingly the bottleneck appears in the processing and interpretation of data.

New methods have been presented here for processing SIMS data. In this chapter a suite of tools for the processing of SIMS data and images in both 2D and 3D is presented. This suite implements solutions for several stages of data processing, including input file reading, peak detection, compression, normalization, visualization, classification and interpretation. The tool-kit implements the methods discussed in previous chapters as well as further extensions which allow for easy data analysis and interpretation. The toolbox has been implemented in Matlab. These tools are to complement already available resources such as those available from the Multivariate Surface Analysis Homepage [39].

5.1.1 Rationale

Processing and visualisation of data is an important step for any analyst. Unfortunately not all instrument software allows the capability of analysing data in a comprehensive way. This is true of current in-house instrumental software. To further process data in addition to acquisition software, Matlab is generally used as the tool of choice. Here analysts have an array of in-built routines to apply to their data. However, not every analyst has the expertise to write or even use scripts successfully in Matlab. One of the goals of implementing this tool-kit was to assist the analyst in processing their complex data.

As has been outlined previously SIMS data can be very large. In Chapter 3 approaches for compression of SIMS images were introduced. Here the goal is to compress the data both in terms of size, for storage or archival purposes , but also for ease of interpretation. Routines such as this only become useful if the data can actually be read into a computer's memory (RAM). In this chapter approaches are presented for reading very large files.

PCA is a central analysis step for many SIMS users (Section 1.3.6). Mean centring of data is common place, however if trying to perform PCA on an image at full resolution some problems arise. The fact that mean centring removes the sparseness of the data is one of these problems as it can cause memory issues in a standard computer. Here an approach is presented to combat this and mean centre SIMS data without the loss of sparseness of the original data.

When performing exploratory analysis of a sample it can be instructive to compare unknown samples with known samples for comparative purposes. To achieve this spectral matching can be applied. By 'matching' an unknown sample with a list or database of known samples and returning a list of chemically similar spectra could help in analysing unknown spectra. In Section 5.2.3 an investigation into some common spectral matching procedures is carried to test their applicability to SIMS data.

Visualisation of 3D image data (actually 4D including the spectral domain) in Matlab can be a task, especially since the data from SIMS machines can be so large and thus need to be read in sparse format. This becomes a problem as Matlab can only deal with 2D sparse data. Two GUIs are presented here for processing data, performing PCA analysis and visualising the results, both for 2D and 3D SIMS image data.

5.2 Data Processing

5.2.1 Importing of data

Images acquired on the J105 can easily reach tens of gigabytes. Reading these files into RAM can be an issue, even considering the sparse nature of the data. In-house routines have been set up to convert the data from the proprietary file format into a format readable by Matlab. The problem then becomes an issue of getting this large data into Matlab memory successfully without exceeding the available memory of the computer.

Routinely analysts bin or use peak selection on acquired data to reduce the size of the data so as to be able to perform MVA in Matlab. This may not be an optimal approach as information is lost when binning or peak selection is used. Some compression techniques were discussed in Chapter 3 to avoid these techniques while achieving large compression. However, some of these routines require the data to be loaded in memory to be performed, while some could conceivably be performed iteratively on single spectra without all the data being present in memory.

Here a workaround is presented for reading very large sparse data files without compression or exceeding memory, given that the size of the data in sparse format does not exceed the memory available. Other approaches to this problem have been addressed by use of encoding schemes [61].

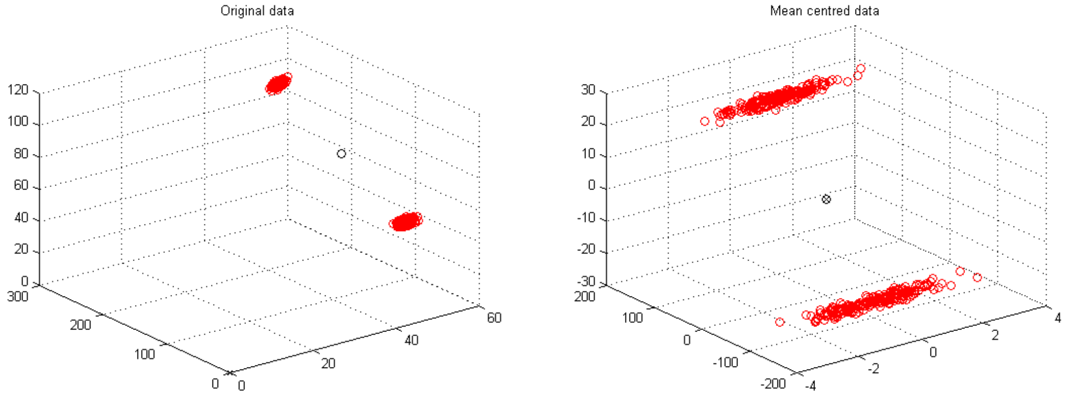
In Matlab to read in sparse data from the disk from an external format the Matlab routine *spconvert* is utilized. A problem can arise with *spconvert* in the way that it reads the data. The routine reads the data as a full matrix and then converts it to a Matlab sparse matrix. The disadvantage of this is that the sparse nature of the data is not conserved throughout the loading process. Thus, this can very easily exceed the memory available on a standard machine. To combat this, a script was designed which breaks the data down into defined chunks. By doing this, the amount of non-sparse data being read can be limited to remain within a feasible limit. After each chunk is read, it is then converted to sparse format. Then all the sparse chunks can be combined together into a sparse matrix containing all the data.

This allows for reading of very large files into Matlab while minimizing the memory requirements. There is however a trade-off between loading time and RAM usage here. Matlab is very fast at matrix operations on sparse data, but can be very slow when trying to do individual operations such as deleting or adding a portion of data. Therefore, the smaller the number of sparse matrix chunks used the faster the loading routine. This means the fewer chunks the quicker the loading, but this also means a greater RAM requirement. This then becomes a trade-off scenario between the two. This setting can be manually altered depending on the machine being used. Reading files in this manner can be useful for other reasons which will be discussed further in the discussion, Section 5.4.

5.2.2 Mean centring sparse data

As has been discussed above, there is a need to keep data in a sparse format to allow it to be stored in the RAM of a standard computer or even of a computer with relatively large RAM (16gb). Storing the non-sparse (dense/full) data has a huge RAM requirement depending on the data being analysed. In general to perform PCA the data are mean centred. This is routinely done so that the ‘axis rotation’ carried out by PCA is through the data from the origin. This however causes a problem for large SIMS data. Since large SIMS data needs to be stored in a sparse format and subtracting the mean from the data causes all the zeros to be changed to non-zero values, thus the sparseness of the data is lost. This then will exceed the memory available within a standard machine if the data are large. Methods for applying memory efficient PCA have been proposed for MS data by [60], however this does not deal with the sparseness issue.

Here a new approach to bypass the memory issues is presented. To illustrate the approach presented, some simple 3D model data were created. Figure 5.1 shows this three dimensional data. The data are separated into two groups and away from the origin, with the mean shown by a black circle. Figure 5.1b shows the same



(a) Model 3D data.

(b) Model 3D data mean centred.

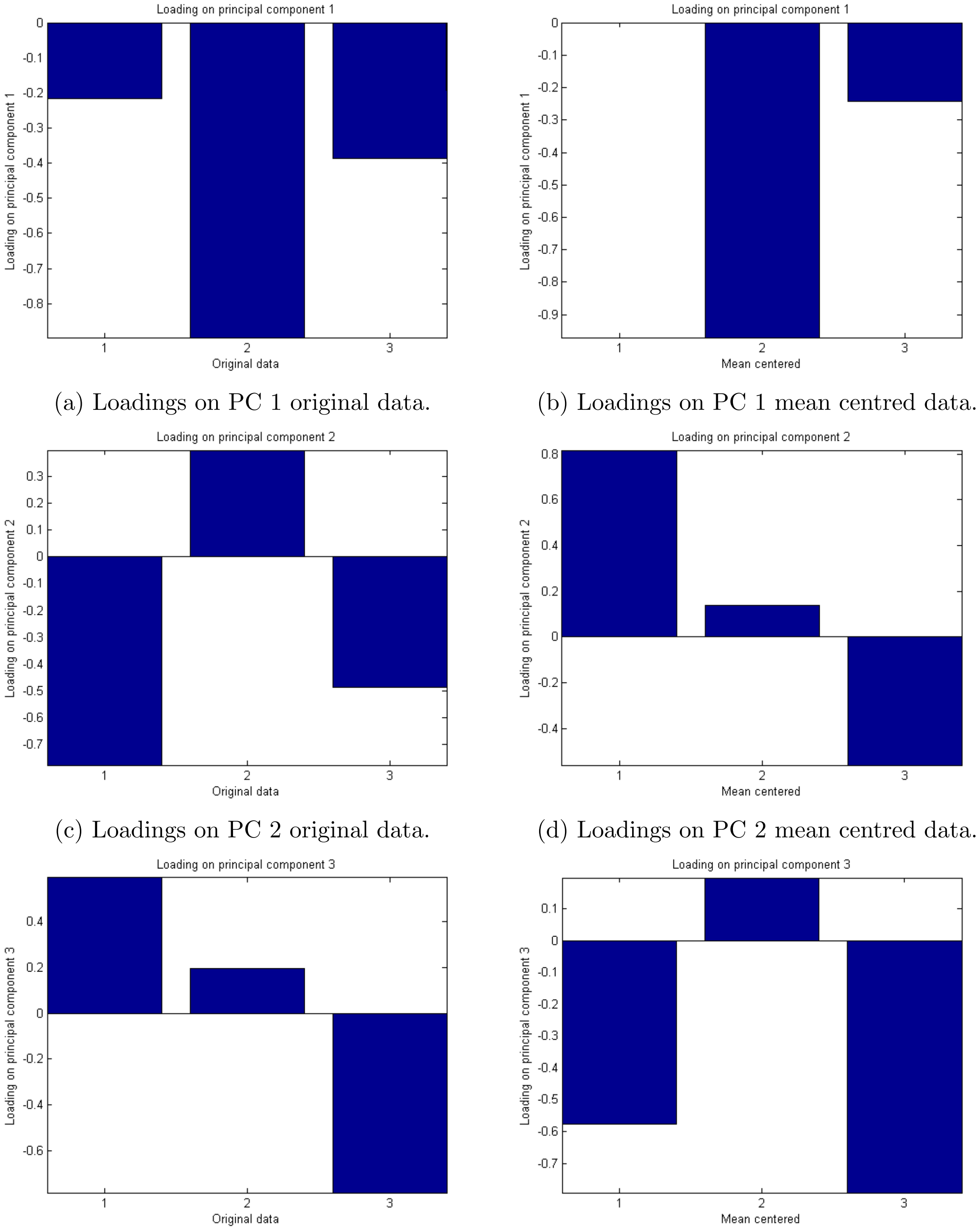
Figure 5.1: Model 3D data before and after mean centring.

data mean centred. To illustrate the rationale for mean centring data for PCA a comparison of the scores and loadings gathered from performing PCA on the data before and after mean centring can be seen in Figures 5.2 and 5.3.

Observing the scores and loadings for the mean centred and original data it is clear there is a large difference between outcomes. Comparing the scores of PC1 *v* PC2 (Figures 5.3a and 5.3b) it is clear that the axis with the largest variance between the data have been selected by the mean centring approach. A full view of the PCs can be seen in Figures 5.3e and 5.3f plotted against their original data. From these it can be seen that mean centring produces a much more optimal approach. With mean the centred data the within group variances can be directly observed from the scores plot of PC2 v PC3, Figure 5.3d. The PCA of the original data, even though there is a separation between the groups, is not as intuitive as with the mean centred data and the within group variance is very difficult to characterise.

In Figures 5.3e and 5.3f the loadings are plotted against their original data. From this it should be noted that the PCA of the original data produces a first PC that is some combination of the mean of the data and the axis with most variance. Whereas the mean centred data has the mean removed and thus has no role in the PCA, therefore the first PC is the axis with the maximum variance. This illustrate why mean centring is important for PCA.

Returning to the point that uncompressed SIMS data cannot be mean centred on a normal computer within the available memory. Thus, another approach that can reach the same outcome as mean centring data is the goal here, without mean centring the original data.



(a) Loadings on PC 1 original data.

(b) Loadings on PC 1 mean centred data.

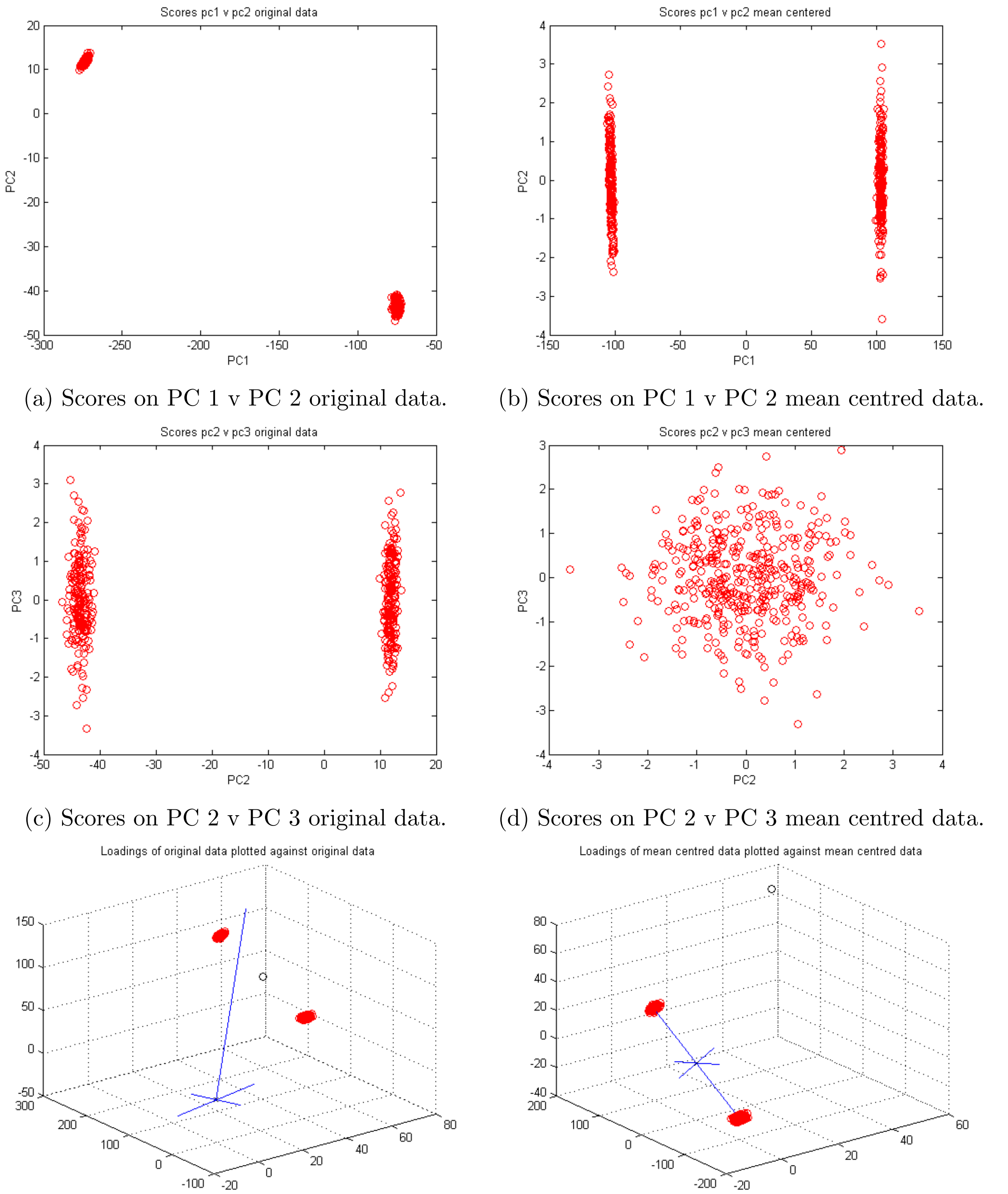
(c) Loadings on PC 2 original data.

(d) Loadings on PC 2 mean centred data.

(e) Loadings on PC 3 original data.

(f) Loadings on PC 3 mean centred data.

Figure 5.2: Loadings of model 3D data before and after mean centring.



(a) Scores on PC 1 v PC 2 original data.

(b) Scores on PC 1 v PC 2 mean centred data.

(c) Scores on PC 2 v PC 3 original data.

(d) Scores on PC 2 v PC 3 mean centred data.

(e) PCA loadings axis plotted against original data.

(f) PCA loadings axis plotted against original mean centred data.

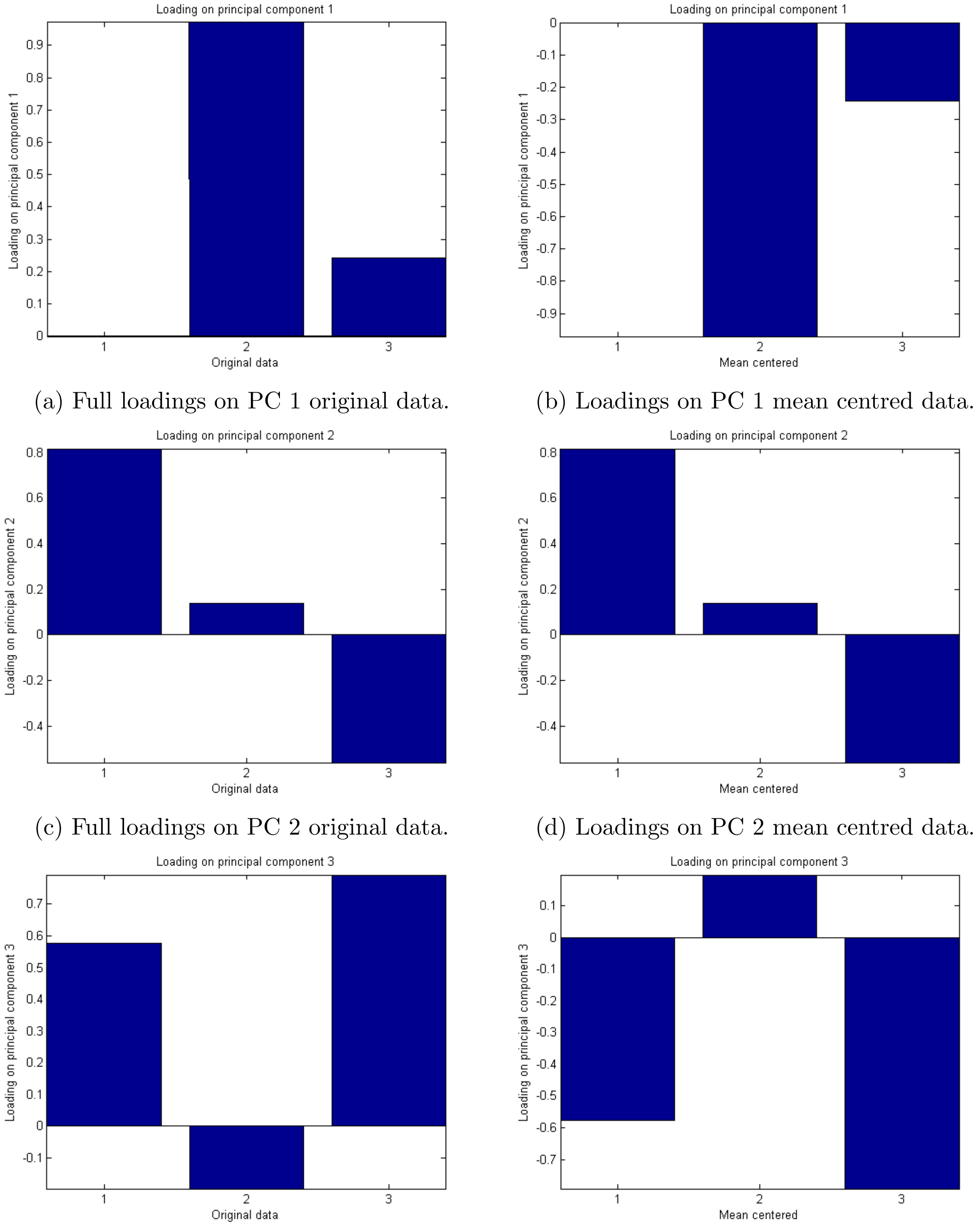
Figure 5.3: Scores of model 3D data before and after mean centring.

Here it is proposed to perform PCA, project the mean of the data into the PC space (or calculate the mean in the PC space, each approach is equivalent), mean centre in the PC space and then redo the PCA on the mean centred scores. This results in a new scores matrix and a new loadings matrix. By multiplying the loadings from both iterations of PCA a 'full' set of loadings can be generated which span the two instances of PCA i.e. the full loadings give the equivalent loadings as if PCA had been applied only once and the final scores are used. Thus, this results in a scores and loadings matrix as if PCA had only been applied once on mean centred data.

At first glance this may seem arbitrary. However, by removing the mean in PC space we can take advantage of PCA's dimensionality reduction. When running PCA most of the variance appears in the first few PCs. By retaining a large number of PCs almost all of the variance can be retained, while achieving a huge compression in terms of the number of dimensions. In this reduced space the mean can easily be subtracted without reaching near the size of the original data. Once the mean has been subtracted in the PC space, PCA can be reapplied to give the same (or similar) results as if the original data had been mean centred.

Figure 5.4 illustrates the approach outlined on the simple 3D data. The left column of figures is the results of new approach. Thus the data are mean centred after the first iteration of PCA and then PCA is reapplied to the scores and the loadings from both iterations are combined. The right column of figure is the results of PCA on the mean centred data as before. On first inspection it may appear that the two results are different as Figures 5.4a and 5.4b are not identical, however PCA is not directionally biased and thus one is simply the opposite direction of the other i.e. they are the same axis. Similarly for Figures 5.4e and 5.4f. This is proof of concept of the new approach.

Figure 5.5a shows the results of the new approach on the original data with the loadings plotted. For comparison the results of PCA on the mean centred data can be seen in Figure 5.5b.



(a) Full loadings on PC 1 original data.

(b) Loadings on PC 1 mean centred data.

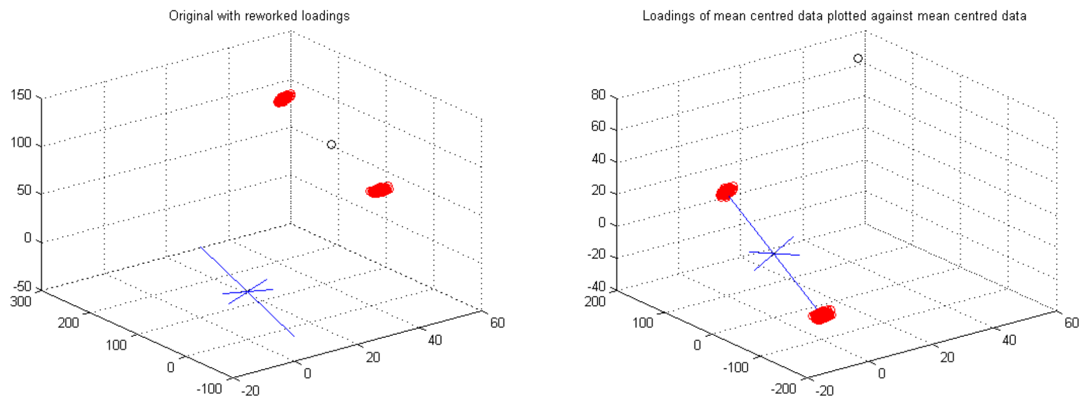
(c) Full loadings on PC 2 original data.

(d) Loadings on PC 2 mean centred data.

(e) Full loadings on PC 3 original data.

(f) Loadings on PC 3 mean centred data.

Figure 5.4: Loadings of model 3D data before and after mean centring.



(a) Model 3D data with PCA loadings plotted of the new approach.

(b) Model 3D data mean centred with PCA loadings plotted.

Figure 5.5: Model 3D data before and after mean centring.

This approach gives exact results for this simple data as it is three dimensional and all the variance was retained throughout, if the dimensionality reduction of the first PCA is to be utilised this does not give the exact results as mean centring the original data. This is because all the variance needs to be captured, meaning a matrix the same size as the original, which is what is desired to avoid.

However, there is a convergence to the correct result depending on the number of PCs used. The more PCs used, the more variance retained, therefore the closer the answer is to the mean centring the original data approach. Table 5.1 list results for some real data acquired on the J105, a section (mass range) of which was used so that it could be mean centred in the available memory to check for convergence. To test how close the new method approximated the classic mean centred results a range of PCs were used. By calculating the distance of each point estimated by the new approach against each point in the classical mean centring approach and summing all these distances, a single value of 'closeness' is calculated, Table 5.1 describes these results. The table are the results from a 100 Da mass range of a 128x128 pixel image. 10 PC's were used for the calculation. Thus the sum of differences result is the sum of the distance of each one of the 16384 pixels in 10 dimensional PC space. From empirical testing 500 PCs used on J105 data gives a

No. of PCs used for initial PCA	Sum of difference from mean centred PCA
100	8194.5
500	988.0821
1000	448.2913
2000	93.0355
4000	.0813

Table 5.1: Results of 'closeness' for different numbers of PC used to estimate mean centred PCA results.

very close approximation.

5.2.3 Spectral Matching

Matching algorithms

There are numerous algorithms for spectral matching or similarity matching which have been applied to spectroscopic data. Three of the most prominent are the dot product, the Euclidean distance and the absolute value distance [68]. These three methods will be tested for their efficacy at matching ToF-SIMS data as in [37]. The accuracy of results depend on various factors. The data must be calibrated correctly, depending on the instrument and the primary ion used, spectra can appear vastly different. For the matching routines to be useful they must overcome these differences to rank the most similar spectra at the top of the 'ranking list'.

Other matching algorithms have also been applied to MS data. Probably the best known of these is the Probability based matching (PBM) system developed by [48]. PBM was developed to identify compounds in mixtures using spectra from GC-MS. PBM has been compared to the three methods tested here and ranked highly [68] but the cosine was found to give the best results. However, in [49] PBM was shown to have a slight advantage. The PBM algorithm is sold commercially, its performance was not tested here.

Other matching routines such as SISCOM [15] used in MassLib, which is software developed to search libraries of mass spectra. SpecInfo is a database which contains mass spectra, IR and NMR data and uses the cosine (or INCOS) method as the

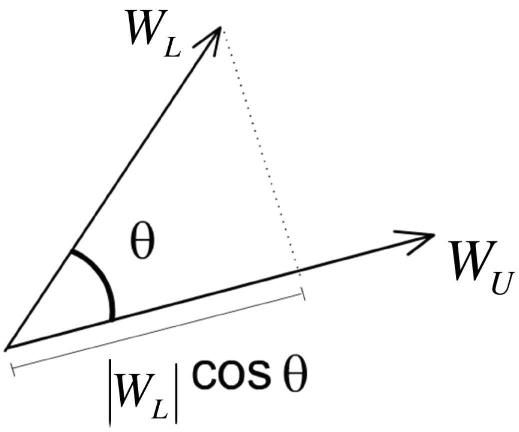


Figure 5.6: Graphical representation of Cosine matching.

search algorithm [52]. Other methods have also been cited [34] [35], however the cosine has performed as well as any of these.

The Dot product

Taking two spectra, one spectrum an unknown and the other a library spectrum, W_L and W_U respectively, and implementing the concept of n-dimensional space the angle between these two vectors can be found. The traditional form of the dot product can be seen in Equation 5.1.

$$W_L.W_U = |W_L|\,|W_U|\,cos(\theta) \tag{5.1}$$

Rearranging Equation 5.1 allows the angle between the vectors to be calculated for an entire mass spectrum; Equation 5.2.

$$cosine = \frac{(\Sigma W_L W_U)^2}{\Sigma W_L^2 \Sigma W_U^2} \tag{5.2}$$

Figure 5.6 shows a graphical representation in two dimensions of two 2D vectors. Using the formula from Equation 5.2 the angle between these vectors in 2 dimensional space and similarly more complex spectra in higher-dimensional space can be found. Since the angle between spectral vectors will always be between 0° and 90°

degrees, as the coordinate for each dimension is along the positive direction, thus the value returned will always be between 0 and 1. A value of 0 means that the angle between the two vectors in n-dimensional space is 90° degrees. A value of 1 indicates that the angle between them in n-dimensional space is 0° degrees, generally neither of these will be reached. Since similar spectra will occupy a similar place in n-dimensional space, the closer the angle between the spectra the more closely related the spectra.

Euclidean distance

The Euclidean distance metric relies on the same principles as the dot product, however the metric evaluates the relationship between two vectors differently. Equation 5.3 shows the formula for the Euclidean distance metric (E).

$$E = \left[1 + \frac{\Sigma(W_L - W_U)^2}{\Sigma W_U} \right]^{-1} \tag{5.3}$$

This is also known as the 'sum of the squares of the differences' between the peak intensities. This metric sums the square difference in each dimension and normalizes over the unknown spectrum. This returns a value between 0 and 1. 0 meaning totally unrelated and 1 meaning exactly the same.

Absolute value

Similar to the Euclidean distance, the absolute value distance uses the same principles but evaluates the vectors differently. Equation 5.4 shows the formula:

$$A = \left[1 + \frac{\Sigma \left| W_L - W_U \right|}{\Sigma W_U} \right]^{-1} \tag{5.4}$$

This is the sum of the absolute differences between the peak intensities. Again this evaluates each corresponding dimension and normalizes over the unknown spectrum. It also returns a value between 0 and 1. 0 being totally dissimilar and 1 being identical.

The latter two metrics (Euclidean and absolute value) can be considered as evaluating spectral points within n-dimensional space. The library spectra create a hyper-volume around the unknown in this space. These metrics evaluate a distance from the unknown point to each point corresponding to the library spectra.

When testing an unknown, a corresponding spectrum/spectra may or may not be contained within the library. Depending on this the outcomes of the search will vary. If the unknown is present within the library, the hyper-volume swept out by the library points in space should contain one or more points which are close to that of the unknown. However, if the unknown is not contained in the library, the hyper-volume should not contain a point close to the unknown. Since there is a lot of variance between even spectra obtained from the same sample, the incorporation of a ranking list is important.

The metrics outlined above generally are only half of a search algorithm. The second half can be considered as data pre-processing.

Data pre-processing

Even though these metrics can return reasonable results, improvements can be made. Pre-processing of spectral intensities can have a beneficial effect on search performance by emphasizing desirable characteristics thus giving more relevant matches. However, the same pre-treatment must be applied to all spectra including the unknown. Data pre-processing can consist of several different approaches from peak selection, probabilistic weighting, peak intensity scaling, and weighting of peak intensity as a function of its mass. Spectra obtained from time of flight mass spectrometers tend to have more numerous large peaks in the lower mass range. This is due to the nature of the SIMS process. As such these low mass peaks hold a large power over the distance metrics. Since higher mass values are often more descriptive it can be instructive to pre-treat the spectra by taking the square root or the cube root of the intensity values (Equation 5.5). This is done because higher mass values tend to have much less intensity. This has the effect of both reducing the relative power of the larger low mass peaks, thus increasing the power of the less intense higher mass peaks.

$$W_{L/U} = \sqrt[n]{I} \tag{5.5}$$

where $W_{L/U}$ is either the unknown W_U or the library spectrum W_L, n is an assigned value and I is the vector of intensity values of that spectrum.

Another form of pre-treatment is by multiplying each intensity by its corresponding m/z value or by its m/z value raised to a power n (Equation 5.6). This again greatly increases the significance of values with a high mass and thus decreases the significance of lower mass peaks.

$$W_{L/U} = I.(m/z)^n \tag{5.6}$$

Other scaling parameters could be applied here, such as those applied in Chapter 4. Results comparing these pre-processing techniques can be found in the Section 5.2.3.

Comparing spectra of various sizes

When acquiring data using a ToF-SIMS instrument, the range of acquisition is set by the user of the instrument. This, while being convenient, means that the ranges that a spectrum is acquired over vary depending on the user. This creates a problem for the spectral matching routines outlined here since they require the input vectors (mass spectra) to be of equal length (dimensions). There are different reasons for this, either the molecule being test had a low mass range and there was no need for a large range acquisition, or the operator did not want or need to see high mass, or the instrument might have been incapable of detecting high masses.

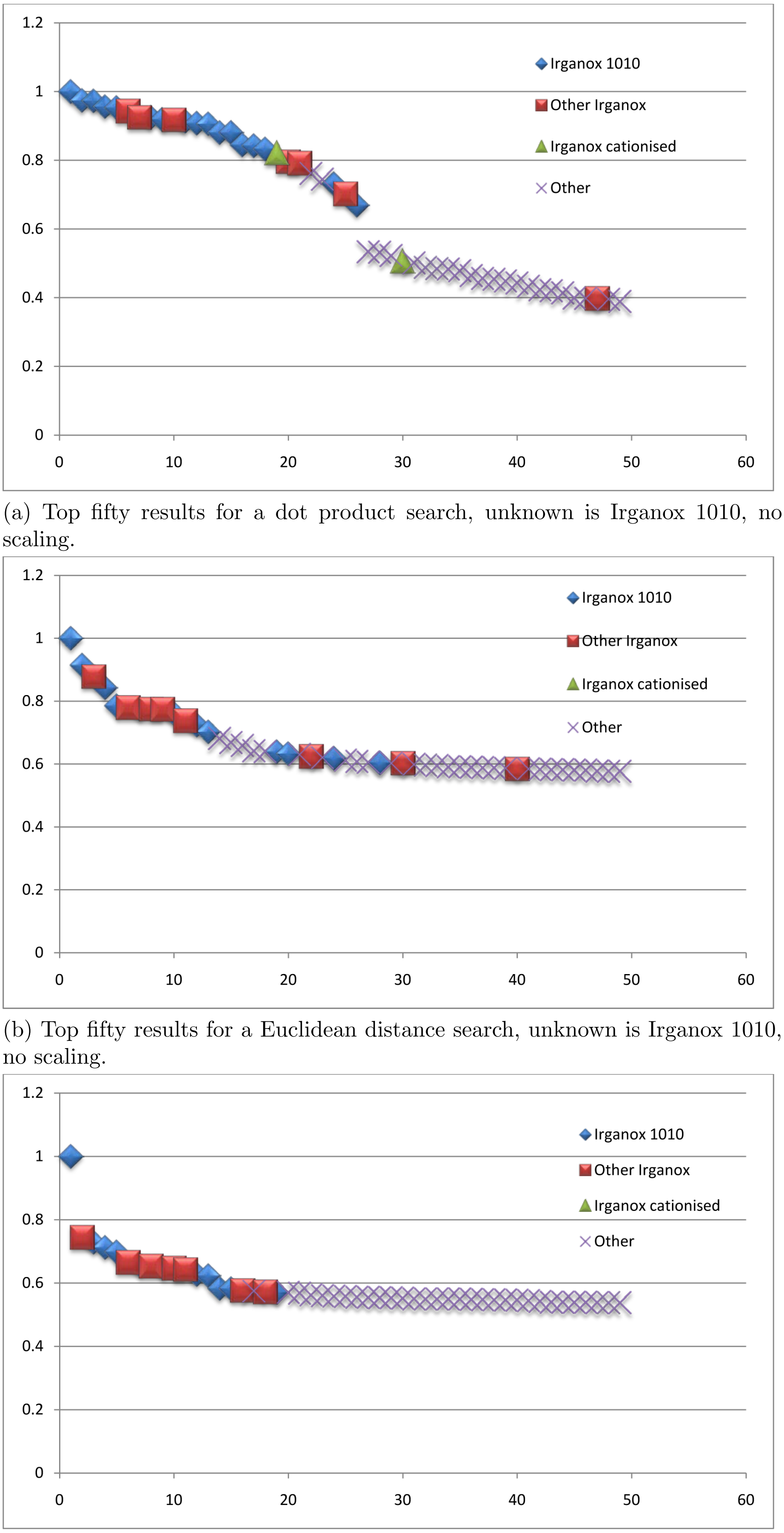
Taking the spectra of different sizes there are two options immediate options; the first is to contract the longer spectra by truncating them, the second is to pad out the shorter spectra with zeros thus increasing their length. For this study both of these options were examined. The rationale here, was to give the unknown spectrum the best fit possible with the known spectra. Therefore, all library spectra were made to the same size as the unknown, thus using all relevant information from the unknown. Padding out shorter spectra may suggest that the best fit is not found. However, padding with zeros does not change the shorter spectrum's point/vector position in n-dimensional space, consequently this will not affect the results of the search. Conversely, if a library spectrum was a match for an unknown but the mass range was too short to incorporate all the relevant data by truncating the larger spectrum the two will have equivalent mass scale and should produce a good match.

Library and test spectra

The SurfaceSpectra Static SIMS Library version 4 is the most comprehensive collection of static SIMS data currently available. It is commercially distributed by SurfaceSpectra ltd. [69]. All matching experiments were carried out on positive ion spectra for which the static SIMS library currently has one thousand. This library, in conjunction with spectra acquired from researchers in the MIB were used in this study. Not all spectra either acquired in-house or in the Static SIMS Library have the same mass ranges or resolution. For this reason all spectra were binned to 1 Da. Though the Static SIMS Library contains numerous different samples it also contains multiple samples of Irganox 1010. This was useful when running these experiments as will be shown in the following sections. Also contained in the library are related molecules of Irganox, such as cationised Irganox 1010 and different structures of Irganox.

Comparison of weighting

Figures 5.7a shows a graph of the top fifty results obtained from matching an Irganox 1010 spectrum contained within the static SIMS library against the other one thousand positive ion spectra. Similarly Figures 5.7b and 5.7c were generated using the same data. These tests were carried out with no scaling applied to the spectra, as



(a) Top fifty results for a dot product search, unknown is Irganox 1010, no scaling.

(b) Top fifty results for a Euclidean distance search, unknown is Irganox 1010, no scaling.

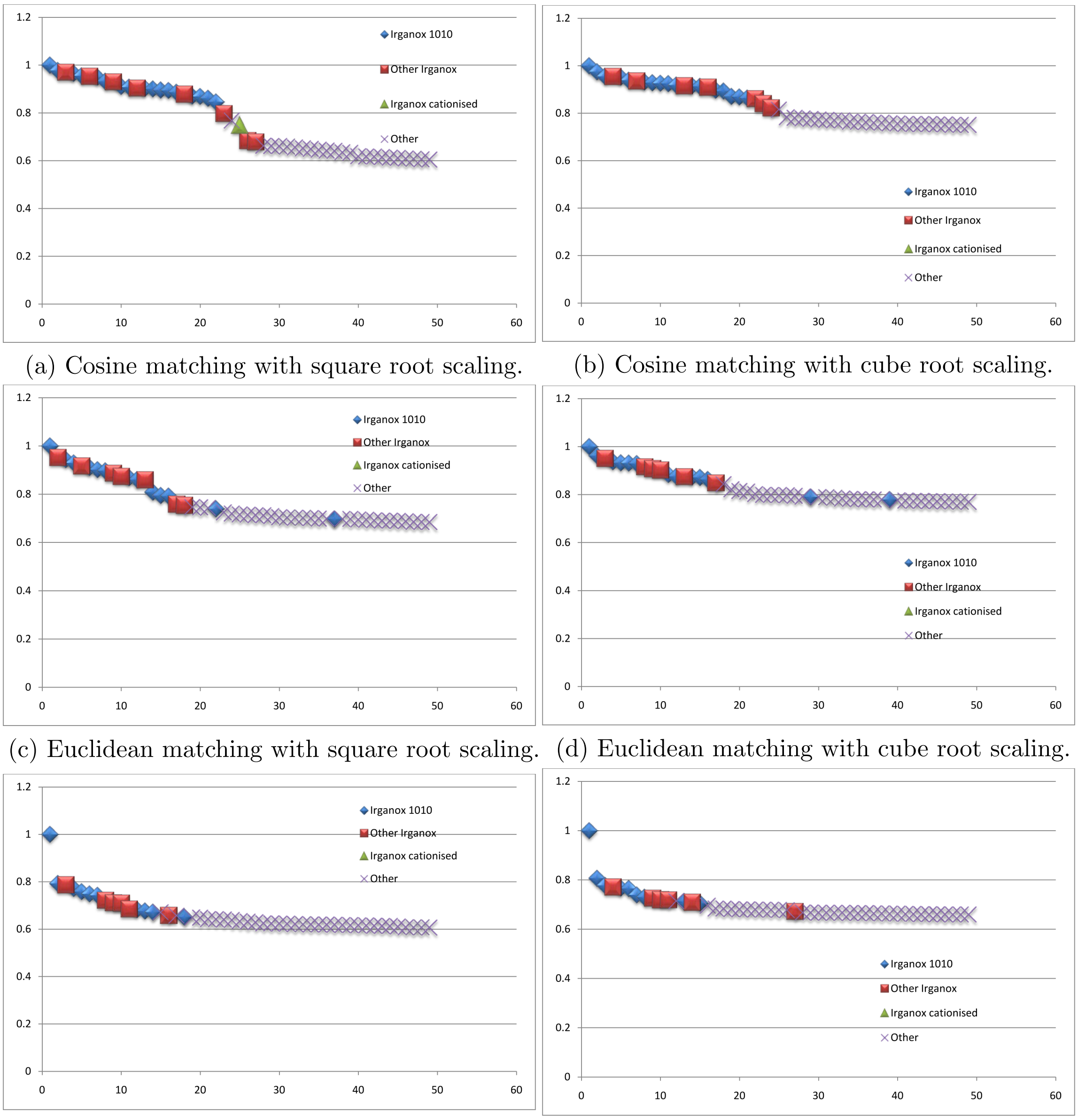
(c) Top fifty results for a absolute value distance search, unknown is Irganox 1010, no scaling..

Figure 5.7: Results of three approaches to spectral matching with no scaling applied.

outlined above (Section 5.2.3).

Comparing each of the approaches against each other it is clear that the dot product metric attains the best results. It performed the best because it obtains the highest proportion of Irganox 1010 matches in the top of the ranking. In comparison with the other two techniques it also has a better range of discrimination, meaning that the value of the results drop away much faster in this technique than with the other two metrics. With this said, a couple of Irganox results which are not 1010 receive high results too. Though not being the same molecule, these do have a similar structure and thus their mass spectra will have many commonalities. This is a good result to see molecules with similar chemistry being ranked highly. This is useful especially if the unknown is not contained in the library. The second best metric appears to be the Euclidean distance. This metric while not performing as well as the dot product, still outperforms the absolute value distance which does not match an Irganox 1010 as its highest rank (disregarding the the first result as this is the test spectrum matched against itself).

Figure 5.8 shows the same search carried out on the same data except with each mass spectrum having been pre-treated by taking the square and cube root of the intensity values respectively. These searches appear to be much less successful than the first, without scaling. However, it has increased the value of the higher matches within the hit list. The order of the ranking has also been change slightly in each of the searches with each benefiting from the scaling in this respect. The cube root scaling appears to be bringing the results closer together, in terms of the value of the result there is less variance across the results, which could lend itself to errors in the ranking and misleading results were there no matching spectrum for the unknown in the library. From reasoning this is what one should expect. When increasing the value by which the root is taken, the intensity values, and hence the variance between spectra will diminish very rapidly. This will in turn make spectra more similar.



(a) Cosine matching with square root scaling.

(b) Cosine matching with cube root scaling.

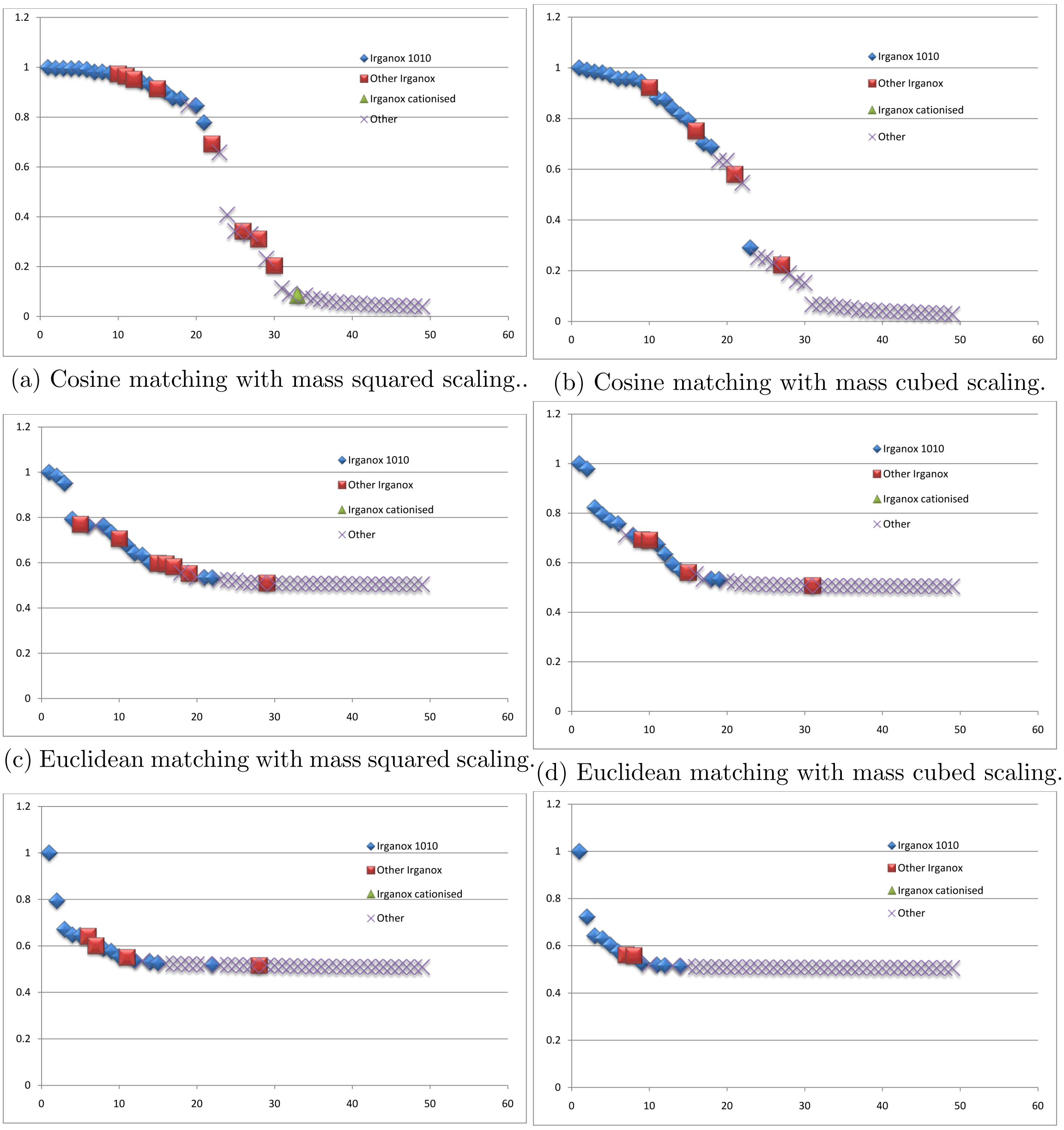
(c) Euclidean matching with square root scaling.

(d) Euclidean matching with cube root scaling.

(e) Absolute matching with square root scaling.

(f) Absolute matching with cube root scaling.

Figure 5.8: Results of matching routines with square and cube root scaling.



(a) Cosine matching with mass squared scaling..

(b) Cosine matching with mass cubed scaling.

(c) Euclidean matching with mass squared scaling.

(d) Euclidean matching with mass cubed scaling.

(e) Absolute matching with mass squared scaling.

(f) Absolute matching with mass cubed scaling.

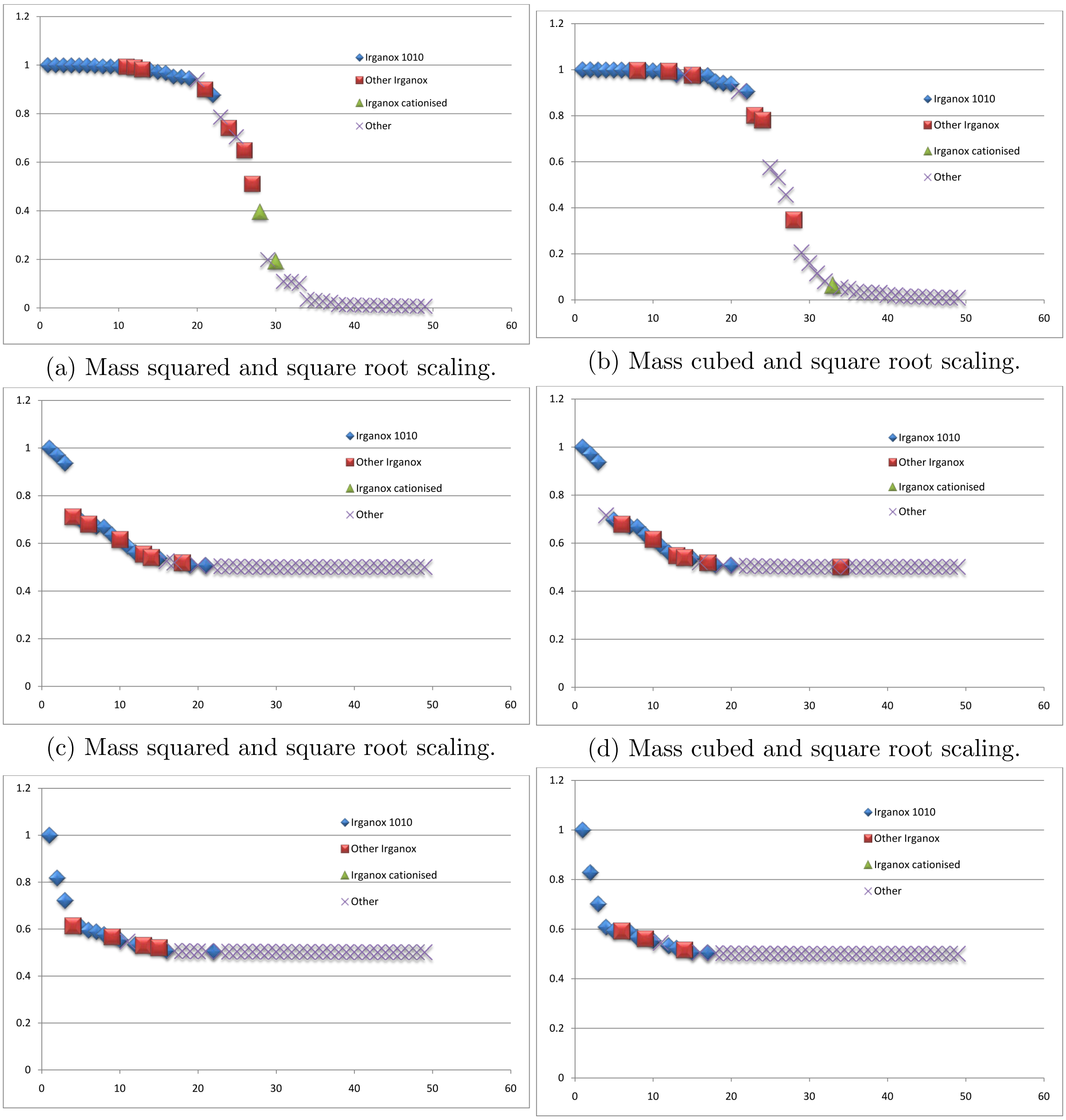
Figure 5.9: Results of matching routines with mass squared and mass cubed scaling.

Figure 5.9 shows the same search with pre-treatment by multiplying each intensity value by the square or cube of their corresponding m/z value. This has a large effect on the outcome of the search, especially for the dot product search. As can be seen the high values in the ranking remain, whereas the lower scored results sharply drop down to a minuscule value (below 10%). This is very advantageous in that it separates high scoring matches from low scoring matches in a decisive way. This, assuming a correct match, makes the ranking much shorter for the purpose of analysis and much easier to interpret. There appears to be little effect on either the Euclidean distance or the absolute value results. However in comparison to the original result, the ranking is slightly better than with no scaling in terms of the population of the top results.

There is no dramatic difference between the square and the cube results for any of the search methods. However, the mass cubed has a higher proportion of desirable results at the top of the ranking. From this, one might propose the product of the mass raised to the power of 4. However, as will be discussed, by increasing this value it increases the probability of false positives, especially with respect to the dot product method.

Figure 5.10 shows the search with all spectra having been pre-treated by both scaling, square root of their intensity and multiplying each intensity by its corresponding m/z value squared and cubed. The cube root scaling will no longer be applied since it reduces the variance of the results by too large a factor.

The dot product results 5.10a and 5.10b clearly depict the most useful set of results. For both scalings, the first 20 results have 16 out of 17 Irganox 1010 at the top with the remaining one just below around position 22. Also three out of the first 20 results are Irganox of similar type but not Irganox 1010. The steep drop from good results to bad result is one of the big advantages of this approach. The ability to discriminate in such a way is a desirable characteristic in a spectral matching routine. In the other two approaches, though correctly match the first couple of results, the proportion to which they rank is not as high as the dot product. Also,



(a) Mass squared and square root scaling.

(b) Mass cubed and square root scaling.

(c) Mass squared and square root scaling.

(d) Mass cubed and square root scaling.

(e) Mass squared and square root scaling.

(f) Mass cubed and square root scaling.

Figure 5.10: Results of matching routines with square root and mass squared/cubed scaling.

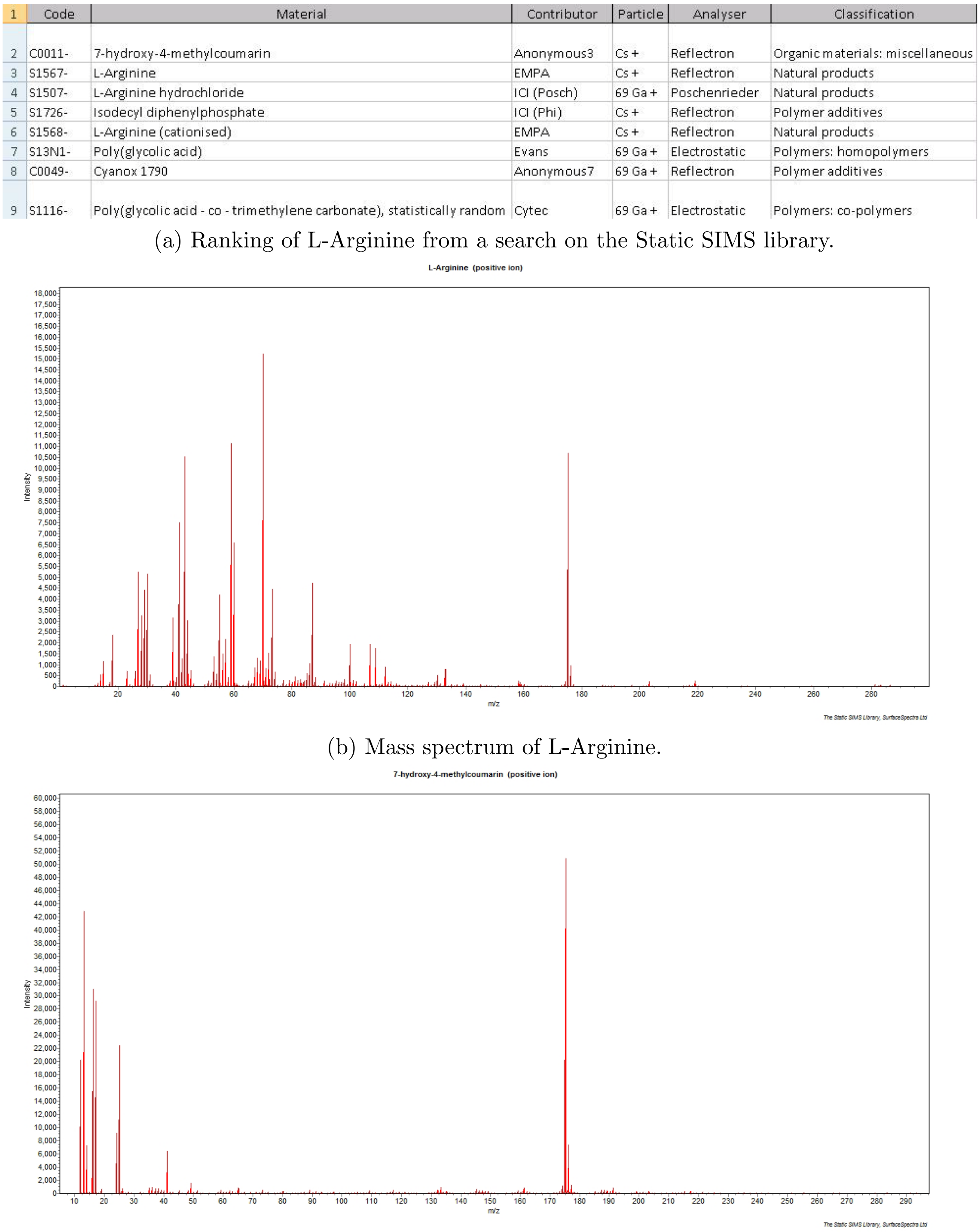
the range over which they do so is much smaller than the dot product and would, had the results not been known, very hard to decipher. From these results it is unambiguous that the dot product method is a much more powerful tool than the other two methods at ranking the similarity of mass spectra.

From the results above it is easy to see the advantages of the pre-treatment schema in regard to the dot product method. Though the other two approaches were improved the results were less evident. Using scaling techniques can ensure a more selective set of results at the top of the ranking list and create a steep drop from high scoring to low scoring results. This has clear advantages for the user as it narrows the field of investigation immensely. There are however some problems with this mode of analysis. One library spectrum has matched very highly in the ranking, especially after scaling. This result outlines one of the problems with the dot product method when scaled as described. This problem and a solution to it will be discussed below.

Problems

Figure 5.11a shows the top eight in the hit list from a search carried out on an L-Arginine sample. The Static SIMS library [69] only contains one spectrum of L-Arginine (it also contains L-Arginine hydrochloride and L-Arginine (cationised) which appear in the ranking), which appears in the second position in the ranking. The top result is a false positive. The root of this problem can be found by inspecting both spectra. In Figures 5.11b and 5.11c the spectra for L-Arginine and 7-hydroxy-4-methylcoumarin can be seen. From observation the most dominant peaks in both spectra can be found at 175Da. This in itself should not be a problem for the distance metrics as many peaks are likely to overlap in any given search and the rest of their intensity peaks are quite different.

However, since both spectra are being pre-treated by taking the square root of intensity and by the product of the intensity and the mass cubed, this places massive influence on this very intense peak. This effect affects the dot product most notably



1	Code	Material	Contributor	Particle	Analyser	Classification
2	C0011-	7-hydroxy-4-methylcoumarin	Anonymous3	Cs +	Reflectron	Organic materials: miscellaneous
3	S1567-	L-Arginine	EMPA	Cs +	Reflectron	Natural products
4	S1507-	L-Arginine hydrochloride	ICI (Posch)	69 Ga +	Poschenrieder	Natural products
5	S1726-	Isodecyl diphenylphosphate	ICI (Phi)	Cs +	Reflectron	Polymer additives
6	S1568-	L-Arginine (cationised)	EMPA	Cs +	Reflectron	Natural products
7	S13N1-	Poly(glycolic acid)	Evans	69 Ga +	Electrostatic	Polymers: homopolymers
8	C0049-	Cyanox 1790	Anonymous7	69 Ga +	Reflectron	Polymer additives
9	S1116-	Poly(glycolic acid - co - trimethylene carbonate), statistically random	Cytec	69 Ga +	Electrostatic	Polymers: co-polymers

(a) Ranking of L-Arginine from a search on the Static SIMS library.

(b) Mass spectrum of L-Arginine.

(c) Mass spectrum of 7-hydroxy-4-methylcoumarin.

Figure 5.11: Results for L-Arginine.

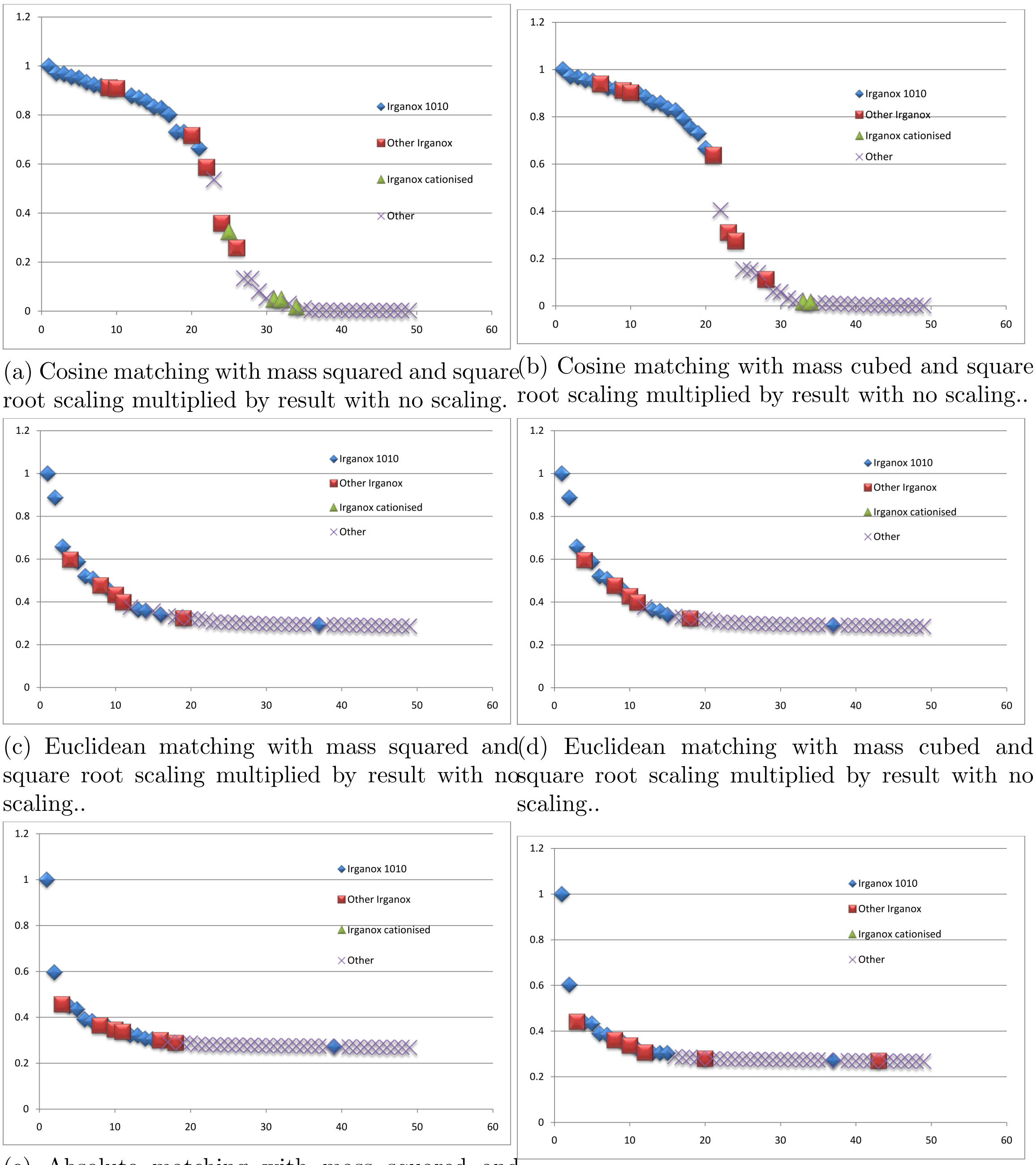
since this peak will be so high in intensity after scaling, both spectral vectors will be pulled largely in this dimension. This has the effect of reducing the angle vastly since the distance travelled in that dimension is so large, hence why the 7-hydroxy-4-methylcoumarin obtains such a high rank. This same effect was also seen in the Irganox 1010 searches where a valine-valine spectrum had an intense peak corresponding to one of Irganox 1010s. This effect is largely due to the pre-treatment of the spectra. However, the same effects are not as apparent in the Euclidean distance or the absolute value as these two methods used the difference between peak intensities to generate their results, therefore the scaling does not affect them in this way. Though this is a feature of the technique it should be apparent to the user when investigating a search why and where this effect can happen. The method should not be treated as a black box method but as a complementary tool.

A suggested solution to this problem is to take the product of the scaled results with that of the unscaled results. By doing this only results that appear high in both rankings will remain high. Thus false positives that arise from the scaling should be minimised. Figure 5.12 show the results of this scaling when applied to each metric.

As was predicted this dropped the values of the false positive from the top of the ranking while keeping good results. This appears to have little effect on the Euclidean distance and the absolute value. The results are not improved and in fact they appear slightly worse than without applying this scaling. For the remainder of the experiments this scaling will be applied to the dot product but not to the other two metrics.

In Table 5.2 the percentile scoring of each method tested in this study are shown. These results were recorded using scaling as outlined above by taking the mass cubed and square root scaling, and the product with no scaling for the dot product. The results are from test run on the Static SIMS library [69] using spectra therein and spectra recorded in house.

From this table it is clear that the dot product method is the most effective matching metric achieving almost 80% correct identification in the top three spectra



(a) Cosine matching with mass squared and square root scaling multiplied by result with no scaling.

(b) Cosine matching with mass cubed and square root scaling multiplied by result with no scaling..

(c) Euclidean matching with mass squared and square root scaling multiplied by result with no scaling..

(d) Euclidean matching with mass cubed and square root scaling multiplied by result with no scaling..

(e) Absolute matching with mass squared and square root scaling multiplied by result with no scaling..

(f) Absolute matching with mass cubed and square root scaling multiplied by result with no scaling..

Figure 5.12: Results of matching routines with square root and mass squared/cubed scaling multiplied by result with no scaling.

	Dot Product	Euclidean Distance	Absolute Distance
First	66.7	53.33	43.33
Top three	10	10	16.67
Outside top three	23.33	36.67	40

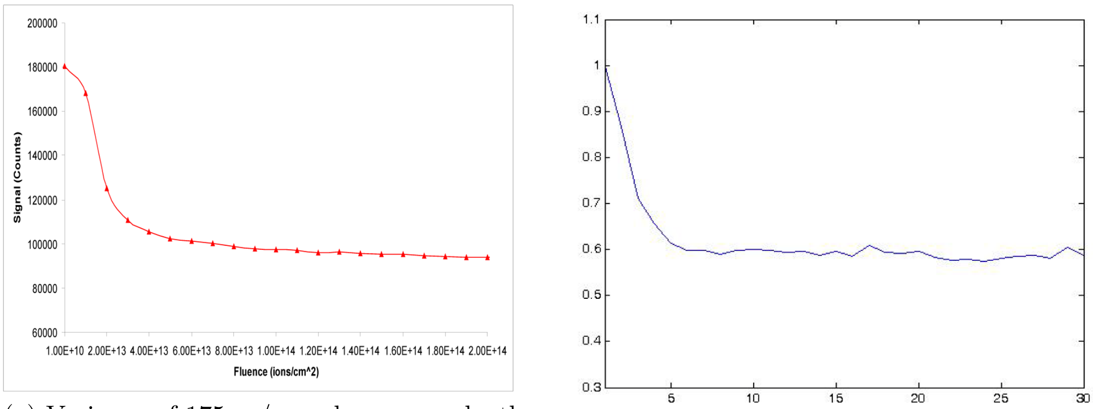
Table 5.2: Percentile scoring of each matching metric as found from testing spectra from the Static SIMS library[69] and in house acquired data against the Static SIMS library.

of the ranking list. This as compared to the Euclidean distance 63.33% and the absolute value distance 60%. These results are in keeping with other studies carried out which test these routines [68]. However, the results found here are lower than those produced elsewhere. This is possibly due to the limited test set of only 30 spectra for this study.

Not described in this results set is the difference seen in the ranking list. However as seen in the results above the dot product was much more effective in this regard. Generally when observing the ranking results, the more a particular result is repeated near the top of the ranking the better, as this should indicate to the user that there is a strong correlation between the two spectra.

Arginine depth profile study

Since the advent of cluster ion beams in ToF-SIMS the concept of 3D imaging has become possible. One of the essential characteristics of cluster beams to be used for this line of analysis is that the damage being caused cannot be more than that which is being removed. What has been found with C_{60}^{+} and other cluster ions is that after an initial period of etching a 'steady state' is reached where the damage caused is essentially equivalent to the damaged material being removed. This can be shown graphically in Figure 5.13a. This graph is the variance of a 175 m/z peak in an L-Arginine sample throughout a depth profile. The 175 m/z peak is the [M+H] ion observed in ToF-SIMS. This means that this peak will characterise the damage being imparted on the sample. This is so because as the damage increases this peak should diminish because damage caused to the molecule will leave fragments and



(a) Variance of 175 m/z peak across a depth profile of L-Arginine.

(b) Result from a dot product search across a depth profile of L-Arginine.

Figure 5.13: Comparison of the variance of the 175 m/z peak of L-Arginine against a spectral match of a L-Arginine depth profile

not the full molecule.

The first point on the graph is an undamaged surface that is being sampled and should be unique to all other data points. Following that, 100% of the surface is calculated to be impacted for each subsequent point. For example, point 2 should have an ion dose of $1.0x10^{13}$ $ions/cm^2$ which is approximately equal to one layer of removal (depending on the sample and definition). Therefore, point three is $2.0x10^{13}$ $ions/cm^2$ and two 'layers' of material should be removed. There is a dynamic that takes place between sample removal and sample damage. On this plot, around point 5 is where sample damage is about equal to damage removal. As can be seen the 'steady state' decreases a little so that there is some extra damage accumulation after point five, but it is minimal compared to the previous four points.

Using the dot product to test a depth profile of L-Arginine, the results in Figure 5.13b were found. It is interesting to see that the variance as found by the dot product across the entire spectrum follows the same trend as that found for the 175 m/z peak. When inspecting the individual spectra in the sets, the trend in the 175 peak can be clearly seen. However, as the 175 peak intensity reduces the fragments ions due to damage increase. This is not as easily tracked by simple observation, but the dot product routine evaluates this extremely well. Similar results were not

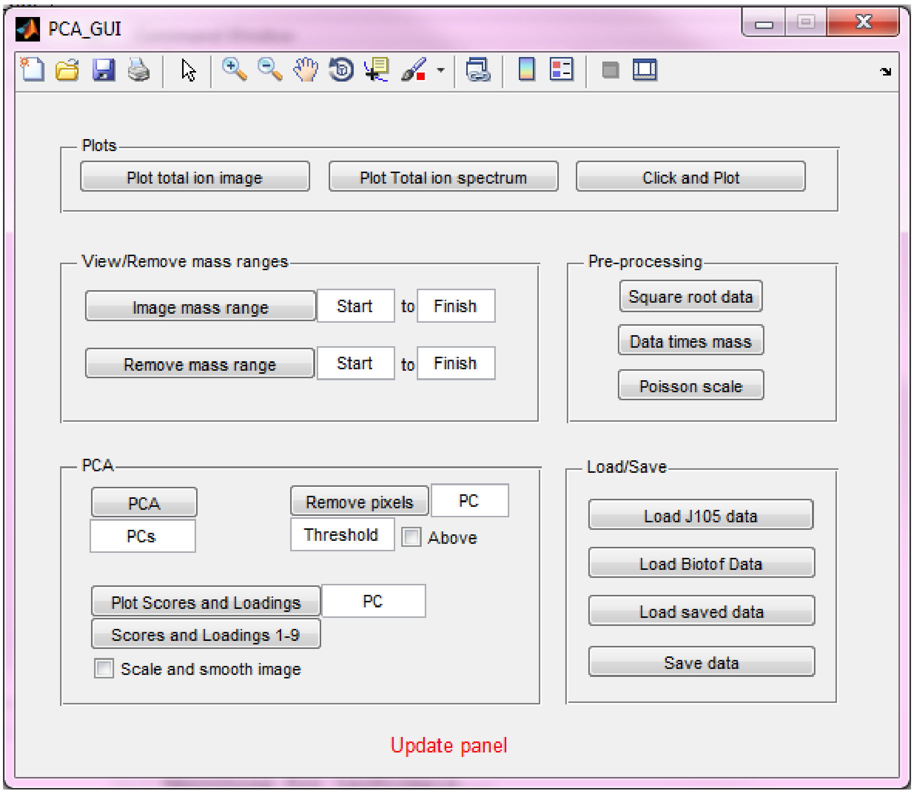


Figure 5.14: PCA GUI.

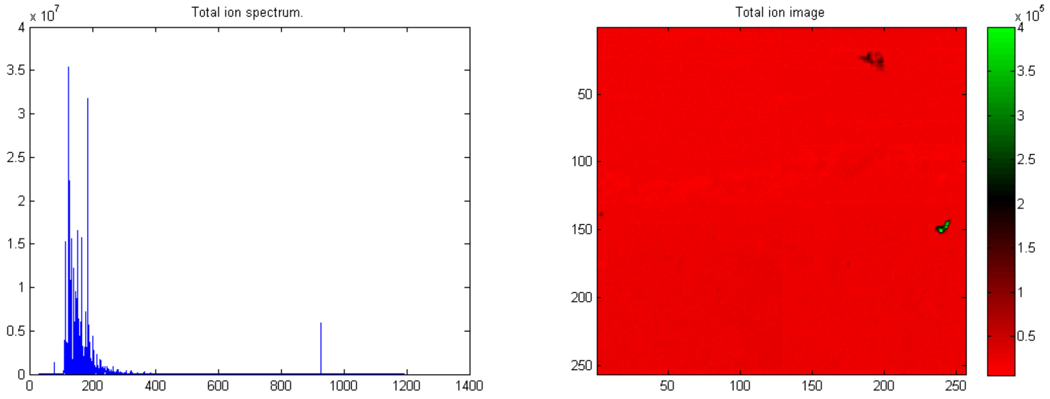
found using the Euclidean distance or the absolute value distance.

5.3 Interpretation toolbox

5.3.1 PCA GUI

Reading files, scaling, processing and visualisation of data in Matlab can be a daunting task for individuals who have little or no experience in programming or indeed Matlab. Here a GUI is presented to read data files and process images acquired either on the Biotof or the J105 data. This GUI can also be easily extended to other file formats. Figure 5.14 shows the GUI for two dimensional images, called PCA_GUI.

This GUI allows for ease of analysis. Simply by clicking on the Load J105 data the analyst can browse their folder structure to the desired image to analyse. The Graphical User Interface (GUI) has been set up with the reading routine for large



(a) Total ion spectrum of sample data.

(b) Total ion image of sample data.

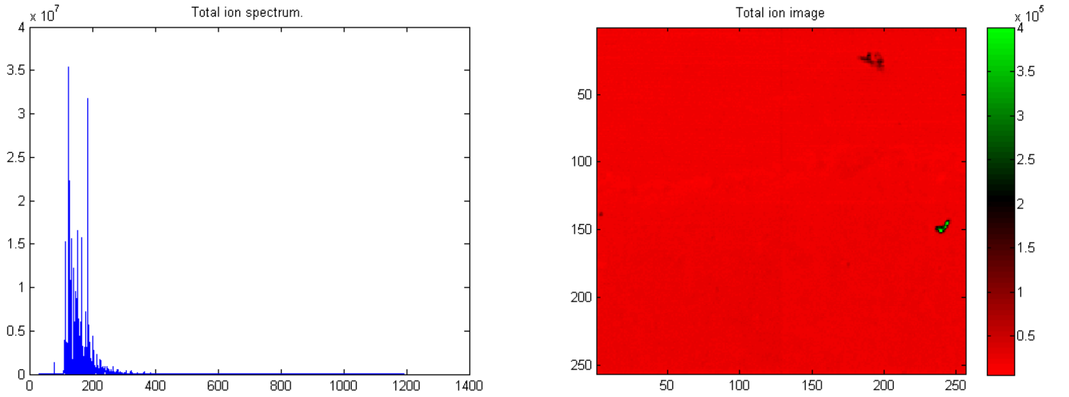
Figure 5.15: Total ion spectrum and image for sample data.

files as described in Section 5.2.1. To fully describe the functionalities of the GUI a sample image is used as an example. The data were acquired on the J105 and is a sample of Prostate tissue, this sample was chosen to exemplify the features of the GUI. Figure 5.15 show the total ion image and the total ion spectrum for the sample.

As can be seen in Figure 5.15b, the image is dominated by some signal arising from two regions (bright green regions). From Figure 5.15a there is a peak which is very intense in comparison to it's surroundings around 900 Da. This peak looks out of place and is possibly a noise peak from switching of the buncher voltage. A noise peak such as this peak was a feature of the J105. To remove this peak the start and ending mass of the desired range to be removed can be input into the remove mass range fields of PCA_GUI.

Figure 5.16 shows the resulting output after removing the noise peak. As can be seen in the new total ion image the high intensity regions still remain and dominate the image. To investigate the source of this high intensity signal one needs to inspect the spectra of these pixels. Using the click and plot function in PCA_GUI this can be achieved.

Click and plot opens a new window (Figure 5.17a) in which the user click on the desired pixels to view the spectrum of this pixel (Figure 5.17b). When the user has



(a) Total ion spectrum of sample data with peak removed.

(b) Total ion image of sample data with peak removed.

Figure 5.16: Total ion spectrum and image for sample data with peak removed.

selected all the pixels they wish to view, they simply hit enter and this ends the click and plot function. This allows the user to inspect the spectra.

Looking at the spectra (Figure 5.17b) from the pixels that dominate the total ion image it is clear that there is very intense signal coming from just below 133 Da. This could be due to caesium contamination. Again the user may wish to remove this signal. By using the functionality of PCA_GUI they can do this.

Figure 5.18 shows the new total ion image after the removal of the 133 Da peak. Now that this intense signal has been removed, a clearer view of the whole sample can be seen. Before proceeding to carrying out PCA on the sample the user may wish to scale the data. Three scaling routines have been inserted in the GUI's functionality. Square root scaling and data multiplied by mass as have been outlined already. Also Poisson scaling has been included for use on data. The description of the PCA section of the GUI will be elaborated on more in the next section where three dimensional data are studied.

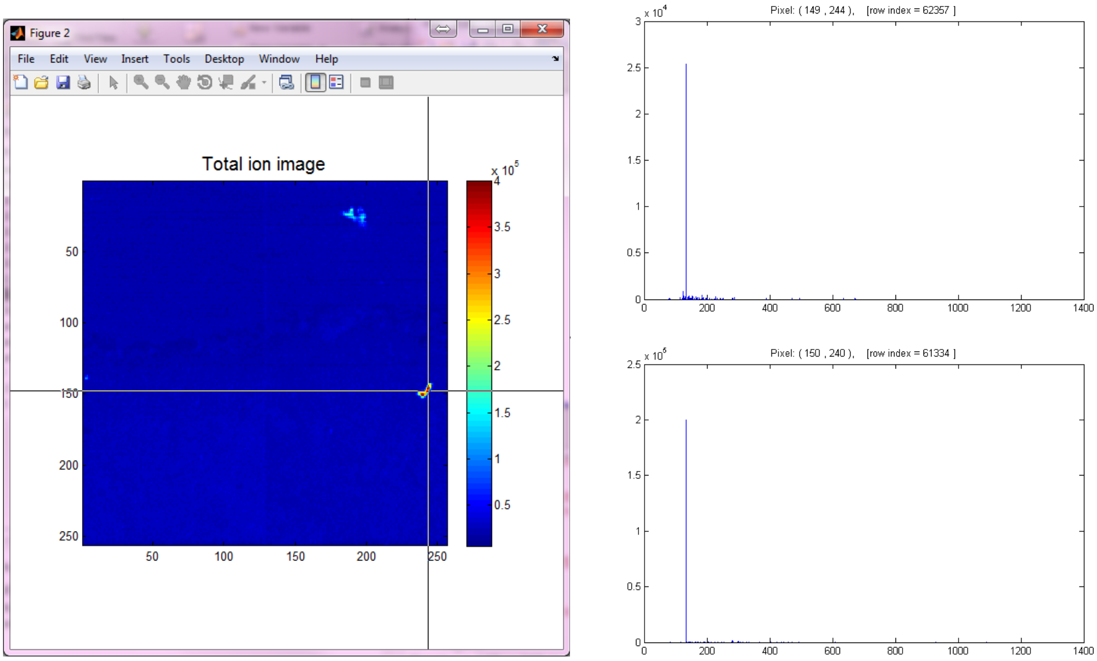


Figure 5.17: The click and plot routine applied to the image data, with two spectra from the high intensity regions.

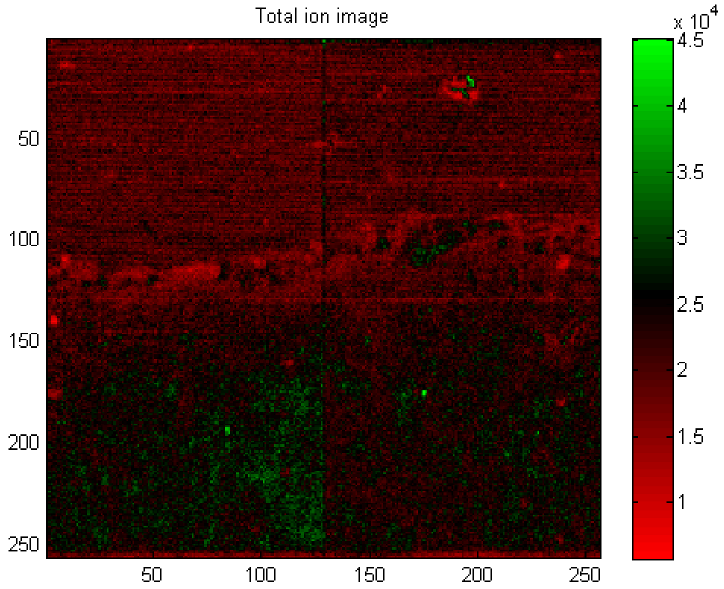


Figure 5.18: total ion image after the intense pixel signal removed.

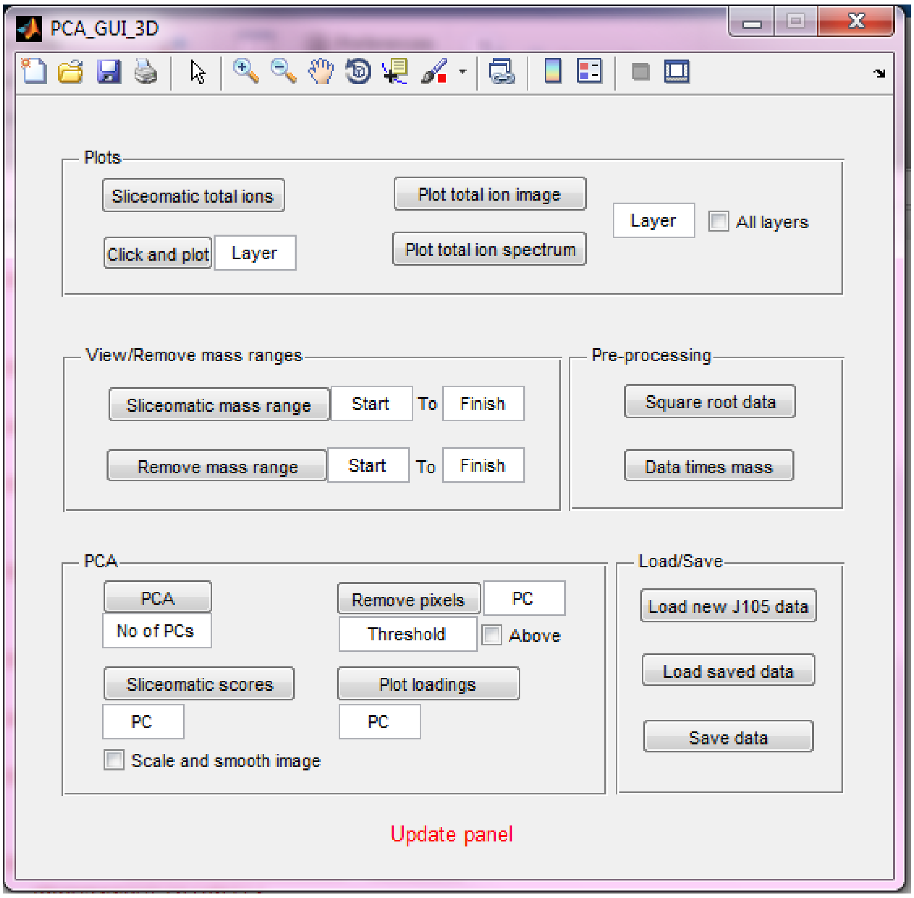
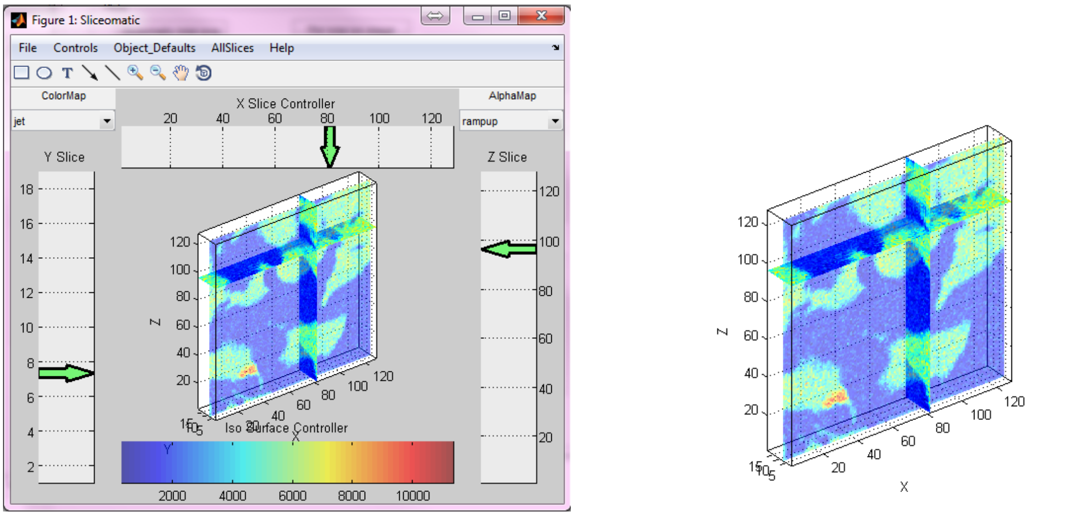


Figure 5.19: PCA_GUI_3D.

5.3.2 PCA_GUI_3D for 3D images

Three dimensional data can now be routinely acquired on instruments equipped with polyatomic sputter sources including the J105. With this, software which is capable of handling such data is also necessary. As has been adressed an approach has been adopted for reading in very large files into Matlab. This approach has again been incorporated in to the PCA_GUI_3D.

Again sample data are used to illustrate the functionality of this GUI. An image stack acquired on the J105 of HeLa-M cells is used [59]. These data are 20 layers of 128 x 128 pixels, with a mass range of 0-600 Da. During acquisition these data were resampled to fixed mass channels of width 0.05 Da. The size of these data on disk is 1.3 gb. Once read into Matlab and written back to disk the data in sparse is only 5 Mb per layer, thus 100 Mb.



(a) Total ion spectrum of real data.

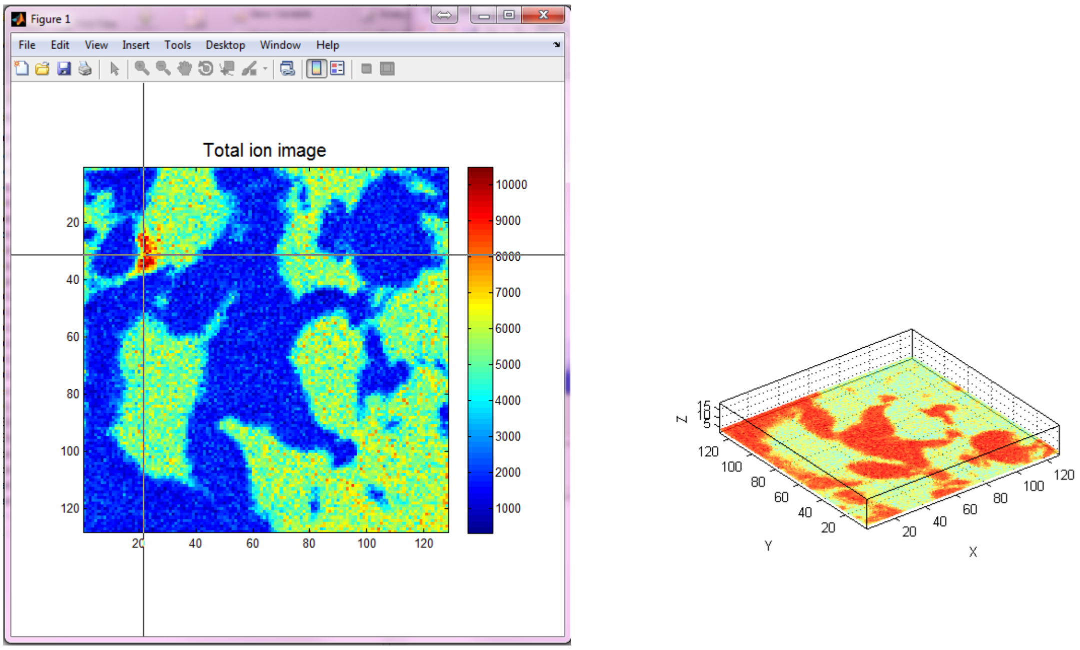
(b) Total ion spectrum of real data.

Figure 5.20: Model 3D data before and after mean centring.

Once the data has been loaded, to keep it in sparse form the matrix must stay in two dimensions as Matlab is incapable of handling three dimensional or four dimensional sparse data. This poses problems in terms of processing and visualisation of the data. Through use of specific calls the data can be visualised without making the whole data ‘full’ or non-sparse. To view the total ion images it is simply the sum of each spectrum at each pixel, thus only one value needed per pixel for visualisation. Similarly for imaging of a mass range. Here the tool Sliceomatic is implemented.

Sliceomatic is the tool used for 3D visualisation of the data. Sliceomatic can be downloaded from [47]. This tool makes visualisation of 3D data easy and intuitive. Figure 5.20a shows the Sliceomatic output of the total ion images of data acquired on the J105. As described these data have 20 layers. By moving the sliders at the top left and right the analyst can view different slices of the data in three dimensions. Figure 5.20 shows an image that has been output from Sliceomatic to be saved or used in report etc. Sliceomatic works on 3D data but as discussed through specific calls to Sliceomatic the 4D SIMS data can be viewed intuitively. By viewing total ion data or imaging mass ranges the SIMS data can be viewed in 3D.

As with the two PCA_GUI, PCA_GUI_3D has the click and plot functionality. The difference here is that a layer needs to be specified to use this function. The



(a) Total ion image for a layer of the image stack.

(b) Scores result from PCA of image stack.

Figure 5.21: Click and plot functionality of image stack 5.21a. 5.21b scores image from PCA of image stack

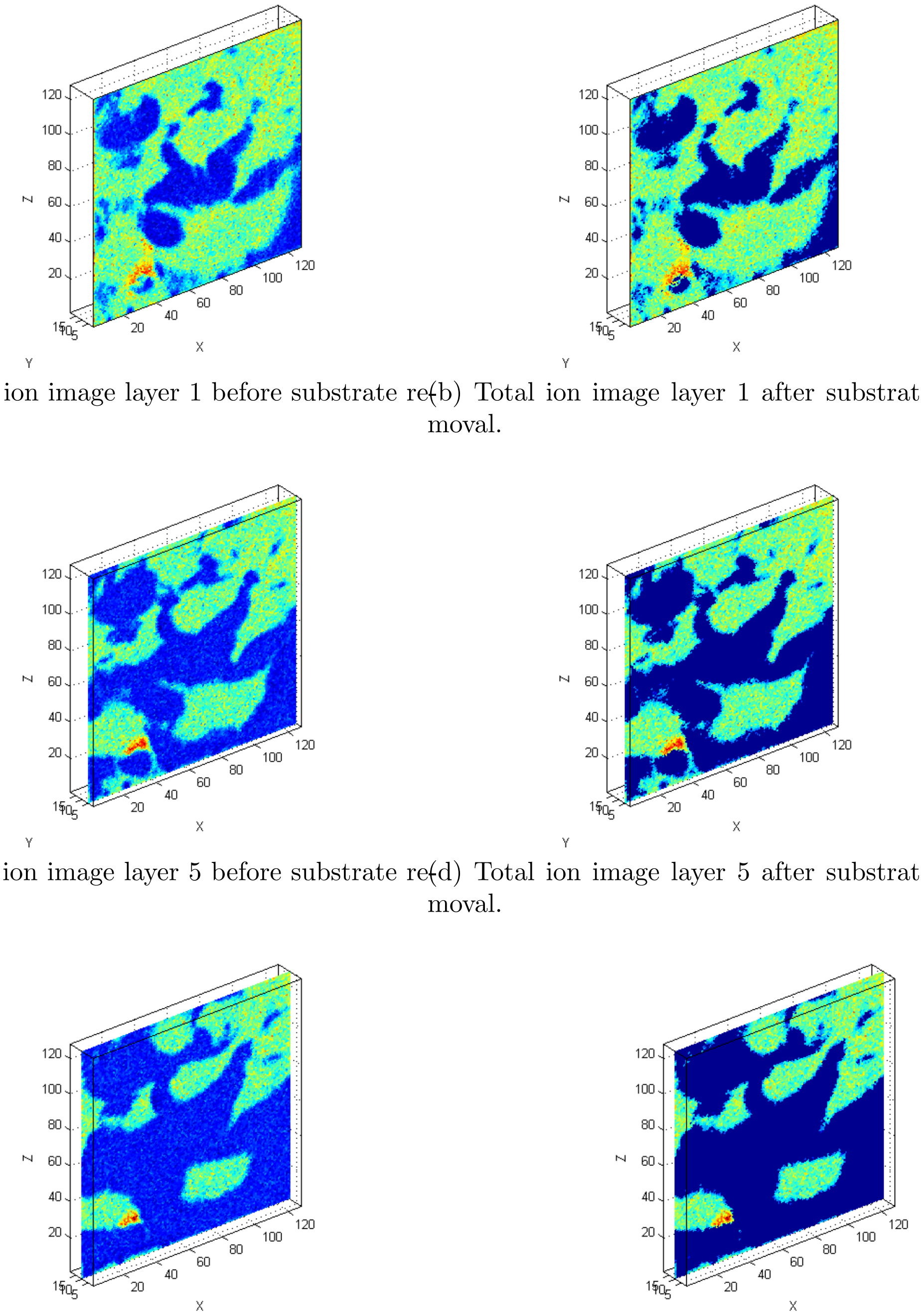
desired layer is input into the GUI. Figure 5.21a shows the click and plot routine active on a layer of this 3D image stack.

Removal of a mass range is also incorporated in the same approach as outlined for PCA_GUI. Similarly, scaling can be directly applied to the two dimensional matrix without losing sparseness. However, scaling routines which seek to mean centre in some way will remove the sparseness of the data and thus can lead to memory issues.

PCA can be directly performed on the sparse two dimensional matrix. The results of this are simply the scores and loadings for the number of principal components, which should cause no memory issues, depending on the number of PCs used. These can be stored in three dimensional matrices as they will be non-sparse data and can easily be viewed, Figure 5.21b.

An additional function is incorporated in the GUI, remove pixels. This function takes in a threshold and a PC number. This function removes pixels from the data that are either above or below a certain threshold within the input PC. This is a quick means of removing substrate or unwanted pixels from an image or image stack.

As can be seen in Figure 5.21b, the substrate pixels have a much different score than pixels arising from the cells. By choosing a value in this PC the substrate pixels can be removed. By doing this further analysis can be performed on the image stack such as PCA on the new data set without the substrate, thus forcing PCA to look at variances within the biological information and not between organic and inorganic. This approach of substrate removal has been used in [25]. Example results of this can be seenin Figure 5.22



(a) Total ion image layer 1 before substrate removal.

(b) Total ion image layer 1 after substrate removal.

(c) Total ion image layer 5 before substrate removal.

(d) Total ion image layer 5 after substrate removal.

(e) Total ion image layer 10 before substrate removal.

(f) Total ion image layer 10 after substrate removal.

Figure 5.22: Total ion images for selected layers before and after substrate removal.

5.4 Discussion

The classification routines mentioned in Chapter 4 were not discussed in this chapter. These are not part of the core toolbox. Though these routines were not written in-house extension to these were written for processing of data. These processing scripts include the leave-one-out cross validation, N-fold cross validation, bootstrapping and scripts for automation of the process. These extension scripts are part of the toolbox and allow for an automated process of classifying data.

All scaling routines referenced previously are available for use. Along with these also the peak picking routine with demos and the compression routines discussed in Chapter 3 are included in the toolbox.

By combining the approaches adopted to read large files, and the approach used to mean centre large sparse data for PCA, the size of data that is able to be read can increase dramatically again. By reading only sections of data and projecting into a reduced PC space this reduces the size of the data in memory vastly and thus larger files can be read. This means that extremely large files, such as big image stack should be no problem to read and perform PCA on without lossy compression of data. A similar approach has been adopted by [55] to read very large files using random projections.

The one important note here is that the PC space to which the data are projected must have most of the variance of the data contained within it. This can easily be done when all the data are present in memory, but difficult to do when reading chunks of data. For example if the first N spectra were read and used to generate the PC space, there is no reason to think that these N spectra are characteristic of the whole image. To subvert this problem we suggest reading selected sections of the data first to obtain the PC space and then reading the rest of the data iteratively and projecting onto this calculated space. To ensure as much of the variance is captured we suggest reading in a spectra from evenly spaced pixels/layers to incorporate as much variance as possible. Every 30th spectrum for example. This approach has not been verified but the author sees no reason why with careful implementation why this should not allow extremely large files to be read in and mean centred PCA performed on them.

5.5 Conclusions

In this chapter a toolbox for analysis of SIMS data has been presented. The toolbox is designed to streamline processing of SIMS data. This includes reading files, scaling, visualising, performing MVA and classifying SIMS data. Approaches for dealing with large data were introduced. Dealing with reading files and a new approach to mean centre sparse data were presented. Results showing the dot product is an effective spectral matching technique for SIMS data were achieved. For this data pre-processing was investigated and was found to have a significant positive effect on the outcome of the search algorithms. Square root of the intensity and the product of the intensity by the cube of its corresponding mass was found to be optimum when taking the product with unscaled results.

Though the contents of the toolbox have been designed and test on in-house data (Biotof and J105), the toolbox is not dependant on these file formats or data types. The two GUIs currently only have support for these data types but could easily be extended to incorporate other file formats. The rest of the toolbox is open for use on any data once the data is read into Matlab.

Chapter 6

Discussion and conclusions

Contents

6.1 Discussion/Future work

This chapter summarizes the thesis by drawing conclusions and discussing possible avenues to follow for continuation of this line of study.

In Chapter 5 a new approach was outlined to mean centre large sparse image data. Some model data were used to demonstrate the rationale for mean centring. However, the model data did not exemplify typical SIMS image characteristics, namely they are hugely sparse. The results of this section clearly showed that mean centring by this approach was possible. It did not however deal with the outcome of PCA on real full resolution highly sparse data. In Figure 6.1 the total ion spectrum of the data used for the study in Chapter 5 can be seen. This is a section of an image acquired on the J105.

These data were used to test mean centring but the results of PCA on the data

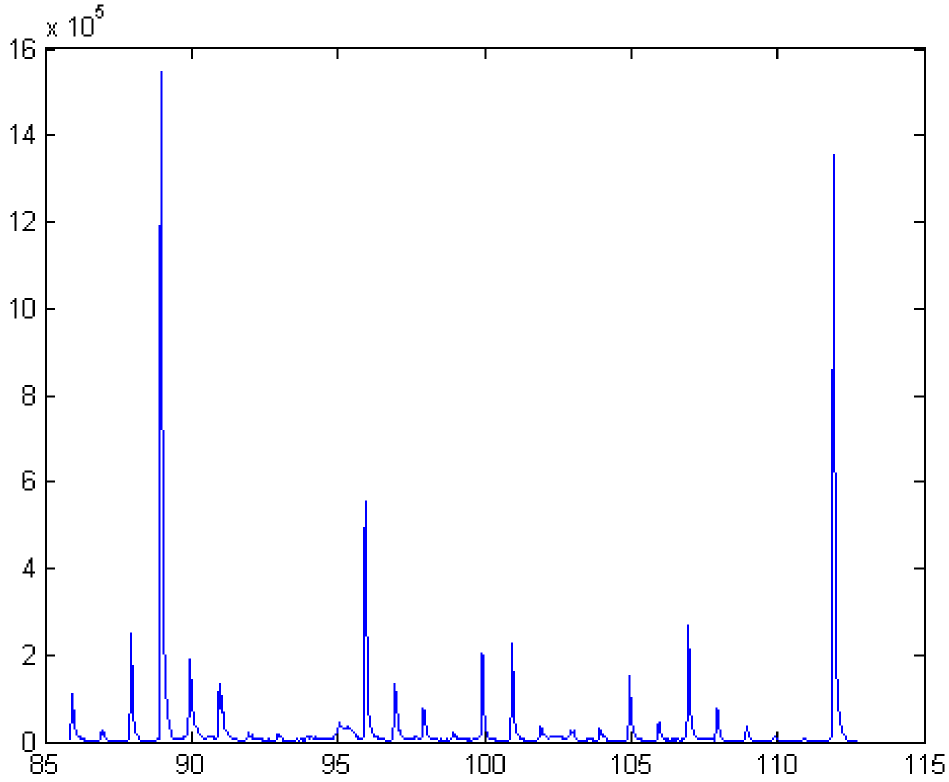
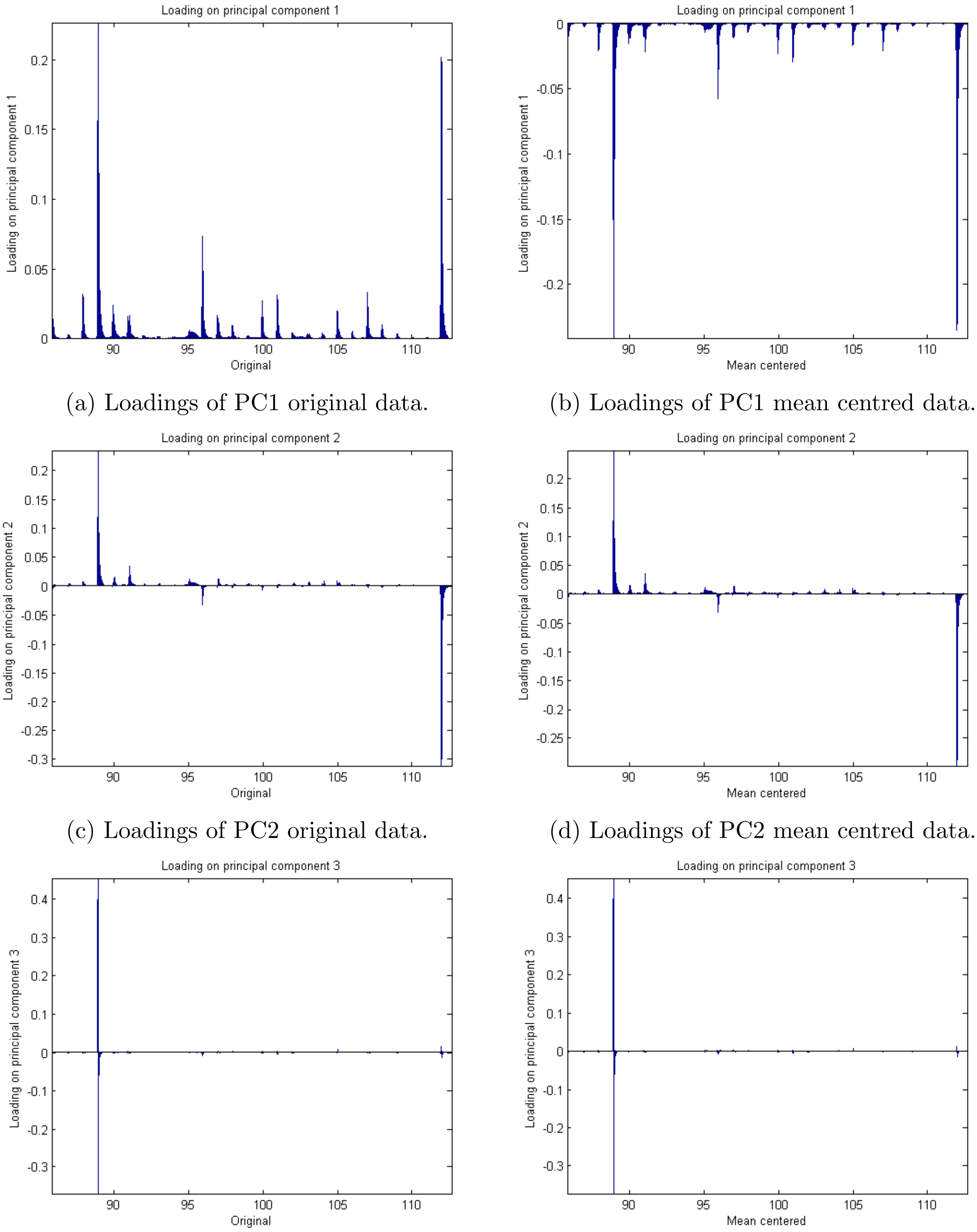


Figure 6.1: Section of real image data.

with and without mean centring were not discussed previously. Figure 6.2 shows the results of PCA, the loadings, with and without mean centring on this data. Examining the loadings it is clear that there is little to no observable difference. Obviously as was calculated in Chapter 5 there are some difference. This raises the question of why the loadings would be so similar when not mean centred.

Consider the mean of this data. Since the data are hugely sparse the mean may well be very small in comparison to signal levels, due to the image being so sparse. Therefore, when mean centring all the zeros move down slightly and all the signal accordingly. When PCA is applied the highest variance in any mass channel is thus going to be the difference between the pixels that have no signal and the pixels that have signal. Even if the mean is not very small, again the largest difference in any mass channel will be the difference between the pixels that have signal and the pixels that do not. This therefore suggests that mean centring of very sparse data such as



(a) Loadings of PC1 original data.

(b) Loadings of PC1 mean centred data.

(c) Loadings of PC2 original data.

(d) Loadings of PC2 mean centred data.

(e) Loadings of PC3 original data.

(f) Loadings of PC3 mean centred data.

Figure 6.2: Loadings of data in Figure 6.1 without (left) and with mean centring (right).

this may be redundant.

As a possible alternative it may be interesting to try mean centring the non-zero data. This would keep all the zero data at the origin for that their respective mass channels and thus they would play no role in the calculation of the PCs. However, the fact that there is no signal in a pixel for a particular mass does not mean there is no information there. The fact that there is no signal is information and a source of variance throughout the image. Thus this approach may produce undesirable results.

The tool-kit presented in Chapter 5 could be extended to include more multivariate approaches such as MAF or MCR. This could easily be done given that the architecture is already in place. Along with other MVA techniques the GUI's could include more of the algorithms outlined in this thesis along with other routines to create a comprehensive analysis tool-kit.

A larger study would be instructive to compound the findings of the spectral matching routines. Also, an ideal identity search algorithm would assign probabilities to the candidate compounds in the hit list that reflect the true likelihood that each is the correct match. This could potentially be done with the current 'score' as assigned by the dot product. However, the limits that would be set for a correct match would be somewhat of an arbitrary empirical threshold. While the methods tested here gave a good indication if the same mass spectrum is in the library being searched, they do not work nearly as well for identifying components in mixtures. By testing patterns such as the distance between peaks more relevant results may be acquired. By applying peak picking algorithms to the data it should reduce the complexity of the data by removing unwanted noise. This may also help the matching metrics improve results.

The classification routines applied here demonstrated the clear ability to perform accurately on SIMS data. To extend the work done here some more exploratory work could be performed to test the ability of the classification approaches to classify across different instruments and ion beams. A possible first step could be to train and test on the data used for the spectral matching study, i.e. Irganox 1010 samples.

The compression techniques outlined in Chapter 3 could be extended by testing their ability to work on image stacks. By investigating incorporating pixels in the neighbourhood region across images as well as within images. Another approach would be to test compressing each individual image in an image stack and recombining the results using alignment algorithms or binning routines. The simplest approach would be to sum all the layers to give an overall total ion spectrum as the starting point and applying the routines to this.

Finally peak picking could be used to automate chemical discovery by locating descriptive peaks in a spectrum and searching chemical databases with relevant information to aid the analyst. This approach overlaps with the work in [37] and could help SIMS in becoming a higher throughput technique.

6.2 Project conclusions

The stated aim at the beginning of this thesis was to improve current methods for dealing with SIMS data. The two broad areas identified in regards to SIMS data were:

- The very large size of current SIMS images/image stacks.
- The complexity of the data.

Large data.

In terms of the size of the data, three approaches were detailed to deal with this problem. Reading sparse data files into memory was the simplest approach, but also an effective workaround. This allows for much larger image data to be read into memory. Following this, compression of images were discussed in Chapter 3 which give a huge reduction in data size without manual peak selection or binning. When this is applied in conjunction with the peak picking routine from Chapter 2 an automated process is created that gives fast, reliable and reproducible results. Thirdly, mean centring of sparse data without memory issues was discussed in Chapter 5. This section illuminates the problem associated with mean centring SIMS image data and negates them by use of PCA's dimensionality reduction. Though this may not be necessary as discussed in Section 6.1.

The complexity of the data.

Peak picking introduced in Chapter 2 is a tool that can be used for compression or as a guide to aid in elucidating spectral chemistry. The results of peak picking can be used in a spectral sense to establish peak lists and for images as an input to compression techniques. In Chapter 5 a tool-kit for importing, processing, visualising and analysing data was detailed. This includes spectral matching for database searching, routines to aid classification and GUI's for automatic processing of data.

In final summary, the routines that have been detailed here were designed to aid SIMS user in the next generation of SIMS instrumentation/experimentation and beyond and to provide deeper insight, accuracy and speed of analysis.

References

[1] Satoka Aoyagi, John S. Fletcher, Sadia Sheraz (Rabbani), Tomoko Kawashima, Irma Berrueta Razo, Alex Henderson, NicholasP. Lockyer, and JohnC. Vickerman. Peptide structural analysis using continuous Ar cluster and C60 ion beams. *Analytical and Bioanalytical Chemistry*, 405(21):6621–6628, 2013. 25

[2] M.J. Baker, E. Gazi, M.D. Brown, N.W. Clarke, J.C. Vickerman, and N.P. Lockyer. ToF-SIMS PC-DFA analysis of prostate cancer cell lines. *Applied Surface Science*, 255(4):1084 – 1087, 2008. Proceedings of the Sixteenth International Conference on Secondary Ion Mass Spectrometry, SIMS XVI. 28, 97

[3] A.R. Bayly, A.R. Waugh, and P. Vohralik. The application of liquid metal ion sources to ion-microprobe secondary ion mass spectroscopy. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 40(5-6):717 – 723, 1985. 24

[4] A. Benninghoven. Z. physik. 230, 403 (1970). 21

[5] Yvon Le Beyec. Cluster impacts at kev and mev energies: Secondary emission phenomena. *International Journal of Mass Spectrometry and Ion Processes*, 174(1-3):101 – 117, 1998. Polyatomic Ion-Surface Interactions. 22, 24

[6] George E. Box and Mervin E. Muller. A Note on the Generation of Random Normal Deviates. *The Annals of Mathematical Statistics*, 29(2):610–611, 1958. 72

[7] Robert M. Braun, Paul Blenkinsopp, Steve J. Mullock, Clive Corlett, Kenneth F. Willey, John C. Vickerman, and Nicholas Winograd. Performance characteristics of a chemical imaging time-of-flight mass spectrometer. *Rapid Communications in Mass Spectrometry*, 12(18):1246–1252, 1998. 104

[8] Richard G. Brereton and Gavin R. Lloyd. Support vector machines for classification and regression. *Analyst*, 135:230–267, 2010. 7, 101, 102

[9] D. Briggs and M.J. Hearn. Analysis of polymer surfaces by SIMS. part 5. the effects of primary ion mass and energy on secondary ion relative intensities.

International Journal of Mass Spectrometry and Ion Processes, 67(1):47 – 56, 1985. 24

[10] Anthony Carado, M. K. Passarelli, Joseph Kozole, J. E. Wingate, Nicholas Winograd, and A. V. Loboda. C60 secondary ion mass spectrometry with a hybrid-quadrupole orthogonal time-of-flight mass spectrometer. *Analytical Chemistry*, 80(21):7921–7929, 2008. PMID: 18844371. 68

[11] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. 101

[12] Catherine Charles, Gervais Leclerc, Jean-Jacques Pireaux, and Jean-Paul Rasson. Introduction to wavelet applications in surface spectroscopies. *Surface and Interface Analysis*, 36(1):49–60, 2004. 40

[13] Kevin R. Coombes, Spiridon Tsavachidis, JeffreyS. Morris, KeithA. Baggerly, Mien-Chie Hung, and HenryM. Kuerer. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, 5(16):4107–4117, 2005. 35, 59

[14] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. 101

[15] H. Damen, D. Henneberg, and B. Weimann. Siscom – a new library search system for mass spectra. *Analytica Chimica Acta*, 103(4):289 – 302, 1978. 126

[16] Soren-Oliver Deininger, Matthias P. Ebert, Arne Futterer, Marc Gerhard, and Christoph Rocken. Maldi imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *Journal of Proteome Research*, 7(12):5230–5236, 2008. PMID: 19367705. 93

[17] Pan Du, Warren A. Kibbe, and Simon M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22(17):2059–2065, 2006. 36, 43, 68

[18] Bradley Efron and Robert Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, 1 edition, 1994. 96

[19] Cecile Engrand, Jochen Kissel, Franz R. Krueger, Philippe Martin, Johan Siln, Laurent Thirkell, Roger Thomas, and Kurt Varmuza. Chemometric evaluation of time-of-flight secondary ion mass spectrometry data of minerals in the frame

of future in situ analyses of cometary material by cosima onboard rosetta. *Rapid Communications in Mass Spectrometry*, 20(8):1361–1368, 2006. 93

[20] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008. 101

[21] J. S. Fletcher, N. P. Lockyer, and J. C. Vickerman. Developments in molecular SIMS depth profiling and 3D imaging of biological systems using polyatomic primary ions. *Mass Spectrometry Reviews*, 2010. 27

[22] John S. Fletcher, Alexander Henderson, Roger M. Jarvis, Nicholas P. Lockyer, John C. Vickerman, and Royston Goodacre. Rapid discrimination of the causal agents of urinary tract infection using ToF-SIMS with chemometric cluster analysis. *Applied Surface Science*, 252(19):6869 – 6874, 2006. Proceedings of the Fifteenth International Conference on Secondary Ion Mass Spectrometry,SIMS XV. 93, 98, 104, 110

[23] John S. Fletcher, Nicholas P. Lockyer, Seetharaman Vaidyanathan, and John C. Vickerman. ToF-SIMS 3D biomolecular imaging of *Xenopus laevis Oocytes* using buckminsterfullerene (C60) primary ions. *Analytical Chemistry*, 79(6):2199–2206, 2007. 5, 26, 27

[24] John S. Fletcher, Sadia Rabbani, Alex Henderson, Paul Blenkinsopp, Steve P. Thompson, Nicholas P. Lockyer, and John C. Vickerman. A new dynamic in mass spectral imaging of single biological cells. *Analytical Chemistry*, 80(23):9058–9064, 2008. 25, 68

[25] John S. Fletcher, Sadia Rabbani, Alex Henderson, Nicholas P. Lockyer, and John C. Vickerman. Three-dimensional mass spectral imaging of hela-m cells sample preparation, data interpretation and visualisation. *Rapid Communications in Mass Spectrometry*, 25(7):925–932, 2011. 151

[26] Max Planck Institute for solid state research. `http://www2.fkf.mpg.de/ga/machines/sims/How_does_TOF_SIMS_work.html`, September 2013. 23

[27] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139, 1997. 98

[28] Yoav Freund and Robert E. Schapire. A short introduction to boosting. In *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1401–1406. Morgan Kaufmann, 1999. 98

[29] D.L. Fugal. *Conceptual wavelets in digital signal processing: an in-depth, practical approach for the non-mathematician.* Space & Signals Technical Pub., 2009. 42

[30] Terrence S. Furey, Nello Cristianini, Nigel Duffy, David W. Bednarski, Michl Schummer, and David Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000. 101

[31] Greg Gillen and Albert Fahey. Secondary ion mass spectrometry using cluster primary ion beams. *Applied Surface Science*, 203-204:209 – 213, 2003. 22

[32] Greg Gillen, Albert Fahey, Matt Wagner, and Christine Mahoney. 3D molecular imaging SIMS. *Applied Surface Science*, 252(19):6537 – 6541, 2006. Proceedings of the Fifteenth International Conference on Secondary Ion Mass Spectrometry. 68

[33] Daniel J. Graham, Matthew S. Wagner, and David G. Castner. Information from complexity: Challenges of ToF-SIMS data interpretation. *Applied Surface Science*, 252(19):6860 – 6868, 2006. Proceedings of the Fifteenth International Conference on Secondary Ion Mass Spectrometry,SIMS XV, Proceedings of the Fifteenth International Conference on Secondary Ion Mass Spectrometry. 28

[34] B. Guzowska-Swider and Z.S. Hippe. Structure elucidation of organic compounds aided by the computer program system scannet. *Journal of Molecular Structure*, 275:225 – 234, 1992. Proceedings of the first National Conference on Molecular Spectroscopy with International Participation. 127

[35] Michael Edberg Hansen and Jrn Smedsgaard. A new matching algorithm for high resolution mass spectra. *Journal of the American Society for Mass Spectrometry*, 15(8):1173 – 1180, 2004. 127

[36] Alex Henderson, John S. Fletcher, and John C. Vickerman. A comparison of pca and maf for ToF-SIMS image interpretation. *Surface and Interface Analysis*, 41(8):666–674, 2009. 70

[37] Alex Henderson, Jimmy D. Moore, and John C. Vickerman. SIMS informatics. *Surface and Interface Analysis*, 45(1):471–474, 2013. 126, 159

[38] Rowland Hill, Paul Blenkinsopp, Stephen Thompson, John Vickerman, and John S. Fletcher. A new time-of-flight SIMS instrument for 3D imaging and analysis. *Surface and Interface Analysis*, 43(1-2):506–509, 2011. 68

[39] Multivariate Surface Analysis Homepage. Multivariate analysis tools for surface analytical data. http://mvsa.nb.uw.edu/, September 2013. 115

[40] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. 2003. 102, 103, 107

[41] ION-TOF. Manufacturers of ToF-SIMS instruments. http://www.ion-tof.com/, September 2013. 27

[42] Michael R. Keenan and Paul G. Kotula. Accounting for poisson noise in the multivariate analysis of ToF-SIMS spectrum images. *Surface and Interface Analysis*, 36(3):203–212, 2004. 25

[43] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. 94, 96

[44] Clemens Lange E FAU Gropl, Knut Gropl C FAU Reinert, Oliver Reinert K FAU Kohlbacher, Andreas Kohlbacher O FAU Hildebrandt, and Hildebrandt A. High-accuracy peak picking of proteomics data using wavelet techniques. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 243–254, 2006. 31

[45] Zhen Li, Stanislav V. Verkhoturov, Jay E. Locklear, and Emile A. Schweikert. Secondary ion mass spectrometry with C60+ and Au4004+ projectiles: Depth and nature of secondary ion emission from multilayer assemblies. *International Journal of Mass Spectrometry*, 269(1-2):112 – 117, 2008. 25

[46] N. P. Lockyer, J. C. Vickerman, and J. S Fletcher. C60, Buckminsterfullerene: its impact on biological ToF-SIMS analysis. *Surface and Interface Analysis*, 2006. 24, 27

[47] Eric Ludlam. Sliceomatic. https://www.mathworks.co.uk/matlabcentral/fileexchange/764-sliceomatic, September 2013. MATLAB Central File Exchange. 149

[48] Fred W. McLafferty and Douglas B. Stauffer. Retrieval and interpretative computer programs for mass spectrometry. *Journal of Chemical Information and Computer Sciences*, 25(3):245–252, 1985. 126

[49] Fred W. McLafferty, Mei-Yi Zhang, Douglas B. Stauffer, and Stanton Y. Loh. Comparison of algorithms and databases for matching unknown mass spectra. *Journal of the American Society for Mass Spectrometry*, 9(1):92 – 95, 1998. 126

[50] Jimmy D. Moore, Alex Henderson, John S. Fletcher, Nicholas P. Lockyer, and John C. Vickerman. Peak picking as a pre-processing technique for imaging time of flight secondary ion mass spectrometry. *Surface and Interface Analysis*, 45(1):461–465, 2013. 69

[51] Jeffrey S. Morris, Kevin R. Coombes, John Koomen, Keith A. Baggerly, and Ryuji Kobayashi. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21(9):1764–1775, 2005. 35

[52] Reinhard Neudert and Michael Penk. Enhanced structure elucidation. *Journal of Chemical Information and Computer Sciences*, 36(2):244–248, 1996. 127

[53] Nha Nguyen, Heng Huang, Soontorn Oraintara, and An Vo. Mass spectrometry data processing using zero-crossing lines in multi-scale of gaussian derivative wavelet. *Bioinformatics*, 26(18):i659–i665, 2010. 35, 40, 43, 68

[54] Nha Nguyen, Heng Huang, Soontorn Oraintara, and An P. N. Vo. Peak detection in mass spectrometry by gabor filters and envelope analysis. *J. Bioinformatics and Computational Biology*, 7(3):547–569, 2009. 31

[55] Andrew D. Palmer, Josephine Bunch, and Iain B. Styles. Randomized approximation methods for the efficient compression and analysis of hyperspectral data. *Analytical Chemistry*, 85(10):5078–5086, 2013. 153

[56] R. Paruch, L. Rzeznik, B. Czerwinski, B. J. Garrison, N. Winograd, and Z. Postawa. Molecular dynamics simulations of sputtering of langmuir-blodgett multilayers by kiloelectronvolt C60 projectiles. *The Journal of Physical Chemistry C*, 113(14):5641–5648, 2009. 24

[57] PHI. Manufacturers of ToF-SIMS instruments. `http://www.phi.com/`, September 2013. 27

[58] Zbigniew Postawa, Bartlomiej Czerwinski, Nicholas Winograd, and Barbara J. Garrison. Microscopic insights into the sputtering of thin organic films on Ag111 induced by C60 and Ga bombardment. *The Journal of Physical Chemistry B*, 109(24):11973–11979, 2005. 24

[59] S. Rabbani, J. S. Fletcher, N. P. Lockyer, and J. C. Vickerman. Exploring subcellular imaging on the buncher-tof J105 3D chemical imager. *Surface and Interface Analysis*, 43(1-2):380–384, 2011. 148

[60] Alan M. Race, Rory T. Steven, Andrew D. Palmer, Iain B. Styles, and Josephine Bunch. Memory efficient principal component analysis for the dimensionality

reduction of large mass spectrometry imaging data sets. *Analytical Chemistry*, 85(6):3071–3078, 2013. 118

[61] Stephen E. Reichenbach, Alex Henderson, Robert Lindquist, and Qingping Tao. Efficient encoding and rapid decoding for interactive visualization of large three-dimensional hyperspectral chemical images. *Rapid Communications in Mass Spectrometry*, 23(9):1229–1233, 2009. 117

[62] O. D. Sanni, M. S. Wagner, D. Briggs, D. G. Castner, and J. C. Vickerman. Classification of adsorbed protein static ToF-SIMS spectra by principal component analysis and neural networks. *Surface and Interface Analysis*, 33(9):715–728, 2002. 93

[63] Abraham. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964. 34

[64] Rob Schapire. Lecture on adaboost. `http://videolectures.net/mlss05us_schapire_b/`, September 2013. 7, 99, 100

[65] M. P. Seah. Analysis of cluster ion sputtering yields: correlation with the thermal spike model and implications for static secondary ion mass spectrometry. *Surface and Interface Analysis*, 39(7), 2007. 22

[66] Levent Sendur and I.W. Selesnick. Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency. *Signal Processing, IEEE Transactions on*, 50(11):2744–2756, 2002. 40

[67] Sadia Sheraz ne Rabbani, Andrew Barber, John S. Fletcher, Nicholas P. Lockyer, and John C. Vickerman. Enhancing secondary ion yields in time of flight-secondary ion mass spectrometry using water cluster primary beams. *Analytical Chemistry*, 85(12):5654–5658, 2013. 25

[68] Stephen E. Stein and Donald R. Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5(9):859 – 866, 1994. 126, 142

[69] SurfaceSpectra. The static SIMS library version 4. `http://www.surfacespectra.com/simslibrary/index.html`, June 2010. 10, 33, 131, 138, 140, 142

[70] C.E. Thompson, J. Ellis, J.S. Fletcher, R. Goodacre, A. Henderson, N.P. Lockyer, and J.C. Vickerman. ToF-SIMS studies of bacillus using multivariate analysis with possible identification and taxonomic applications. *Applied Surface*

Science, 252(19):6719 – 6722, 2006. Proceedings of the Fifteenth International Conference on Secondary Ion Mass Spectrometry,SIMS XV. 28, 97

[71] J.J. Thompson. Rays of positive electricity, 1913. 19

[72] Hua Tian, Stephen Riechenbach, Qingping Tao, and Alex Henderson. Classification and cluster analysis of complex time-of-flight secondary ion mass spectrometry for biological samples. *2009 International Conference on Bioninformatics, Computational Biology, Genomics and Chemoinformatics*, 2009. 93, 104

[73] John F. J. Todd. Recommendations for nomenclature and symbolism for mass spectroscopy. *International Journal of Mass Spectrometry and Ion Processes*, 142(3):209 – 240, 1995. 20

[74] B. Tyler. Interpretation of ToF-SIMS images: multivariate and univariate approaches to image de-noising, image segmentation and compound identification. *Applied Surface Science*, 203204(0):825 – 831, 2003. Secondary ion mass spectrometry SIMS XIII. 31

[75] Bonnie J. Tyler. Multivariate statistical image processing for molecular specific imaging in organic and bio-systems. *Applied Surface Science*, 252(19):6875 – 6882, 2006. Proceedings of the Fifteenth International Conference on Secondary Ion Mass Spectrometry,SIMS Proceedings of the Fifteenth International Conference on Secondary Ion Mass Spectrometry. 28

[76] Bonnie J. Tyler and Richard E. Peterson. Dead-time correction for time-of-flight secondary-ion mass spectral images: a critical issue in multivariate image analysis. *Surface and Interface Analysis*, 45(1):475–478, 2013. 25

[77] Bonnie J. Tyler, Gaurav Rayal, and David G. Castner. Multivariate analysis strategies for processing ToF-SIMS images of biomaterials. *Biomaterials*, 28(15):2412 – 2423, 2007. Imaging Techniques for Biomaterials Characterization. 28

[78] J. C. Vickerman and D. Briggs. *ToF-SIMS: Surface Analysis by Mass Spectrometry.* SurfaceSpectra/IMPublications, 2001. 21

[79] John C. Vickerman. *Surface Analysis - The Principal Techniques.* Wiley, July 1997. 22, 23

[80] M.S. Wagner, T.A. Horbett, and David G. Castner. Characterizing multicomponent adsorbed protein films using electron spectroscopy for chemical analysis, time-of-flight secondary ion mass spectrometry, and radiolabeling: capabilities and limitations. *Biomaterials*, 24(11):1897 – 1908, 2003. 31, 68

[81] D. E. Weibel, N. Lockyer, and J. C. Vickerman. C60 cluster ion bombardment of organic surfaces. *Applied Surface Science*, 231-232:146 – 152, 2004. Proceedings of the Fourteenth International Conference on Secondary IOn Mass Spectrometry and Related Topics. 25

[82] Baolin Wu, Tom Abbott, David Fishman, Walter McMurray, Gil Mor, Kathryn Stone, David Ward, Kenneth Williams, and Hongyu Zhao. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19(13):1636–1643, 2003. 101

[83] Chao Yang, Zengyou He, and Weichuan Yu. Comparison of public peak detection algorithms for maldi mass spectrometry data analysis. *BMC Bioinformatics*, 10(1):4, 2009. 31, 33, 36, 52

[84] Yutaka Yasui, Margaret Pepe, Mary Lou Thompson, BaoLing Adam, George L. Wright, Yinsheng Qu, John D. Potter, Marcy Winget, Mark Thornquist, and Ziding Feng. A dataanalytic strategy for protein biomarker discovery: profiling of highdimensional proteomic data for cancer detection. *Biostatistics*, 4(3):449–463, 2003. 32, 35

[85] J Zhang, E Gonzalez, T Hestilow, W Haskins, and Y Huang. Review of peak detection algorithms in liquid-chromatography-mass spectrometry. *Current Genomics*, 10(6):4, 2009. 33