University of Birmingham

PhD Thesis

Information Processing for Mass Spectrometry Imaging

Author:
Andrew Palmer

Supervisors:
Dr. Josephine Bunch
Dr. Iain Styles

A thesis submitted to the University of Birmingham for the degree of DOCTOR OF PHILOSOPHY

School of Chemistry
School of Computer Science
College of Engineering and Physical Sciences
University of Birmingham
2014

Acknowledgements

Thanks guys!

There are many people that deserve thanks for supporting me throughout my graduate studies, on both an academic and personal level.

The academic supervision I received from Dr. Iain Styles and Dr. Josephine Bunch provided guidance and support throughout. You gave me enough freedom to make research exciting whilst keeping the projects on track. Your commitment to your research and your students has been an inspiration. Without your vision and willingness to collaborate I would not have entered the field of mass spectrometry imaging or spectral image analysis. I have received endless patience and support from my partner Rachel Brown during my Ph.D studies and for this you receive my heartfelt thanks.

Thanks go to the members of the research groups that I have had the pleasure of working alongside. Bunch group past and present should be commended for their consistently professional attitudes and the high quality science they produced. Within the office the tea and coffee was always flowing. Particular thanks go to Rory Steven and Alan Race for many evening of beer and science and Rian Griffiths for office biscuits and always booking the highest quality conference accommodation. Thanks also go to Medical Imaging Research Group in Computer Science, in particular Jamie Guggenheim and Hector Basevi who have provided great companionship during the hours spent in the darkroom. There are too many members of the doctoral training centre to thanks you individually but your friendship helped to keep morale high.

Without funding I could not have embarked on this thesis so sincere thanks go to the EPSRC and the Physical Sciences of Imaging Doctoral Training Centre for supporting me financially.

Abstract

Mass Spectrometry Imaging (MSI) is a sensitive analytical tool for detecting and spatially localising many thousands of ions generated across intact tissue samples. Ionisation techniques that produce intact biomolecules directly from tissue samples, such as Matrix Assisted Laser Desorption Ionisation (MALDI), allow the exploration of sample biochemistry. The datasets produced by MSI are large both in the number of measurements collected and the total data volume, which effectively prohibits manual analysis and interpretation. However, these datasets can provide insights into tissue composition and variation, and can help identify markers of health and disease, so the development of computational methods are required to aid their interpretation.

To address the challenges of high dimensional data, randomised methods were explored for making data analysis tractable without completely discarding any portion of the data. Random projections provided over 90% dimensionality reduction of MALDI MSI datasets, making them amenable to visualisation by image segmentation. Randomised basis construction was used to construct an approximate basis for the data so it could be compressed to 1% of its original size and decompressed with a Signal to Noise Ratio (SNR) of > 40. Automated data analysis was developed that could be applied to compressed data, including segmentation and factorisation, providing a direct route to the analysis and interpretation of MSI datasets. Evaluation of these methods alongside more established dimensionality reduction pipelines was performed on simulated and real-world datasets and they were shown to efficiently extract the spatial patterns present.

Randomised methods were found to provide a powerful set of tools for applying automated analysis to MSI datasets, which would otherwise have been impossible. Whilst developed specifically on MALDI MSI these methods have been extended to other hyperspectral imaging modalities. They provide a route for sharing datasets between researchers for validation and comparison of processing algorithms whilst simultaneously accelerating the performance of existing approaches. They provide researchers with a tool for investigating the increasingly large datasets that are produced by hyperspectral imaging in applications such as high-resolution and 3D studies.

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

ADC Analogue to Digital Converter

BASC Basis Approximation for Spectral Compression

CHCA α-cyano4-hydroxycininnamic acid

DESI Desorption Electo-Spray Ionisation

H&E Haematoxylin and Eosinn

HWHM Half Width Half Maximum

FFPE Formalin Fixed Paraffin Embedded

FWHM Full Width Half Maximum

ICA Independent Component Analysis

IRF Instrument Response Function

JL Johnson-Lindenstrauss

LC Liquid Chromatography

MALDI Matrix Assisted Laser Desorption Ionisation

MCP Micro Channel Plate

MS Mass Spectrometry

MSI Mass Spectrometry Imaging

m/z mass to charge ratio

NASH Non-alcoholic Steatohepatitis

NNMF Non-negative matrix factorisation

PC Phosphatidylcholine

PCA Principal Component Analysis

PCC Pearson's Correlation Coefficient

RGB Red Green Blue

ROC Receiver Operating Characteristics

ROI Region of Interest

RMS Root Mean Square

SNR Signal to Noise Ratio

SOM Self Organising Map

SIMS Secondary Ion Mass Spectrometry

SVD Singular Value Decomposition

TDC time to digital converter

TIC Total Ion Chromatogram

TFA TriFluoroacetic Acid

TOF Time of Flight

Chapter 1

Introduction to Mass Spectrometry Imaging (MSI) Acquisition, Analysis and Interpretation

---

This chapter provides a background to the instrumentation of untargeted MSI of biological samples and the methods for processing the data. It summarises the features that can be visible in the data and how acquisition parameters effect the final spectra. A review of algorithms developed for automatic determination of molecular distributions and spectral profiles is provided which highlights the difficulties encountered in identifying molecules with important distributions from within the large datasets produced by MSI.

1.1 Overview

Ionisation techniques that produce intact biomolecules directly from tissue samples, such as Matrix Assisted Laser Desorption Ionisation (MALDI), allow the direct exploration of sample biochemistry, with a particularly noteworthy application being thin tissue sections. MSI uses this sensitive analytical tool for de-

tecting many thousands of ions at known spatial locations across a sample. The datasets collected consist of tens-of-thousands of spectra each with hundreds-of-thousands to millions of mass to charge ratio (m/z) measurements. In an exploratory (or untargeted) experiment there may be ions within this large volume of data that have well defined spatial distributions corresponding to the sample biochemistry. So, these datasets provide a way to simultaneously probe multiple biochemical micro-environments within tissue, providing insights into molecular variation and help to identify markers of health and disease.

To make this information accessible to researchers it is necessary to extract a smaller number of images that depict the sample heterogeneity and show which spectral patterns are associated with them. However, the size of the data is a major roadblock in identifying patterns within the data, the datasets produced by MSI are large both in the number of m/z measurements collected (very high dimensional data) and the total number of spectra (making the data volume large). The high dimensionality prohibits manually sifting through all the possible ion images as this is impossibly time consuming so computation methods are required to make their interpretation possible. Unfortunately, the high dimensionality also restricts the application of automated computational approaches.

Many popular machine learning techniques for automated pattern detection, such as segmentation or factorisation, have been applied to MSI datasets and shown to extract biologically distinct patterns of ions. All required the dimensionality of the data to be reduced beforehand, principally to limit the data size (overcoming computational limitations) but also to concentrate measurements (e.g. by discarding noise or combining identical m/z patterns) which makes such algorithms more effective. However, MSI datasets are often so large that the dimensionality reduction algorithms cannot be applied to them due to computational restrictions and some data must first be discarded. Random projection is a data-independent dimensionality reduction approach that is both computationally and memory efficient so that it can be applied directly to the large datasets produced by MSI. This provides a novel route to analysing the patterns within a whole MSI dataset without explicitly discarding any portion of the data.

The aim of this work is to implement and evaluate randomised methods for addressing the high dimensionality of real-world and simulated MSI datasets providing a computationally efficient route. This approach differs quite radically from established dimensionality methods and so metrics and methods for comparison across approaches need to be developed. Realistically simulated data provides opportunities for making quantitative comparisons between algorithms (and the opportunity for comparisons between laboratories). The validated approaches can then be applied to data collected from several real-world sources to demonstrate the application of the methods developed.

1.1.1 Organisation of this Thesis

The remainder of this chapter provides a survey of the literature and reviews the acquisition of exploratory mass spectrometry images and common strategies for data mining and describes the sources of the main problems faced during the processing of MSI data. There is a particular focus on factors that affect the composition of 'untargeted' spectra, methods for extracting and presenting information that is biologically pertinent, how these methods can be evaluated and how this impacts the relevance of MSI to wider biological research fields. Chapter 2 develops the application of random projections for the analysis of MSI and extends this to produce a compression scheme based on randomly detected data patterns. These methods are also shown to provide routes for sharing data as well as retaining the properties of MSI data required for automated analysis (whilst compressed). Chapter 3 explores several automated data analysis techniques for determining molecular profiles from MSI data, all of which would be impractical or impossible to apply without the use of randomised methods. To understand the properties of the dimensionality reduction methods Chapter 4 contrasts several schemes from the literature against a baseline provided by randomised methods. To address to limits of comparative analysis a mass spectrometry imaging data simulator is developed in Chapter 5 and used to evaluated spectral clustering, a popular clustering approach that has thus far not been applied within MSI. Chapter 6 applies a selection of the methods developed within the thesis to aid in the analysis of data acquired from human and animal samples to demonstrate the use of these new pipelines for approaching biological problems.

1.1.2 Publications from this thesis

Peer Reviewed Journal Papers

- The use of Random Projections for dimensionality reduction for segmentation of the liver data in Chapter 6 has been published as: AD Palmer, J Bunch, IB Styles, The Use of Random Projections for the Analysis of MALDI Mass Spectrometry Imaging Data, Journal of The American Society for Mass Spectrometry, 2014.

- Portions of Chapters 2 & 3 on basis approximation have been published as: AD Palmer, J Bunch, IB Styles, Randomised Approximation Methods for the Efficient Compression and Analysis of Hyperspectral Data, Analytical Chemistry, 2013.

- The porcine eye data from Chapter 6 has been published as: Sucrose cryoprotection facilitates imaging of whole eye sections by MALDI mass spectrometry AD Palmer, R Griffiths, I Styles, E Claridge, A

Calcagni, J Bunch Journal of Mass Spectrometry 2012.

Conference Papers

- The application of basis approximation methods to Raman Microscopy were presented at the European Conferences on Biomedical Optics 2013 as: Faster tissue interface analysis from Raman microscopy images using compressed factorisation.

Conference Talks

- The application of basis approximation methods to Raman Microscopy were presented at the European Conferences on Biomedical Optics 2013 as Faster tissue interface analysis from Raman microscopy images using compressed factorisation
- The application of basis approximation methods to 3D MALDI MSI was presented at Mass Spectrometry Applications in the Clinical Laboratory 2014 as Molecular analysis in 3D using imaging mass spectrometry and randomised compression methods.

1.2 Exploratory MSI

Mass spectrometry is a powerful analytical technique which separates gas phase ions based on their relative masses and charge. A portion of molecules from a sample are ionised and then each of the ions are very precisely weighted and counted. The ion abundance is plotted as a function of their m/z giving a spectrum with a peak in signal intensity for each detected ion. The m/z of a particular peak is determined by the elemental composition of an ion which is approximately the sum of the mass of the constituent atoms. In this way a spectrum can be related back to the chemical composition of a sample and this provides the fundamental contrast mechanism for MS. The cohort of peaks within a spectrum provide a molecular profile determined by the chemical composition of a sample. The technique of MSI uses mass spectrometry to collect molecular profiles directly from a sample's surface whilst maintaining knowledge of where ionisation occurred so that spatio-temporal snapshots of molecular composition can be recorded[200]. Since the first demonstration of MALDI MSI for producing spatially resolved distributions of individual molecules from biological samples[31, 200] the technology and methods have advanced massively and within a single experiment images of thousands of molecular ions can be produced simultaneously using an exploratory imaging approach.

The ability to visualise molecules in an multiplexed manner provides a particularly powerful tool for biological samples as traditional histology is only able to visualise a few molecules simultaneously[221]. As

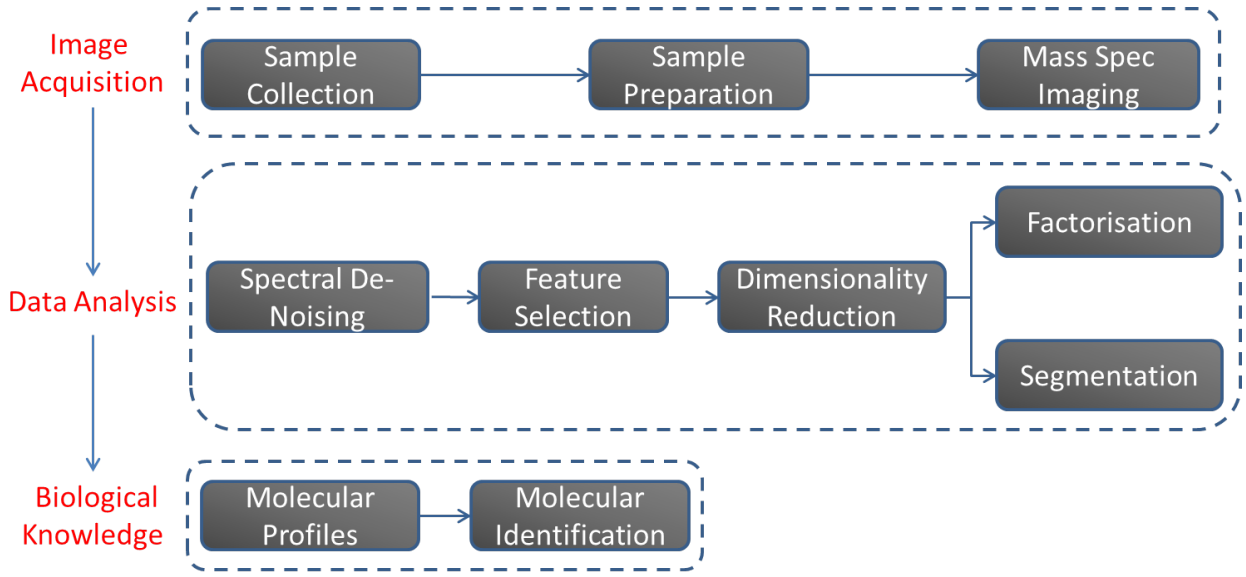

Figure 1.1: Workflow of a MSI experiment with exploratory data analysis (segmentation and factorisation) illustrating some subdivisions of the key experimental and interpretation stages

signals from all ionised molecules are recorded these datasets are rich in chemical information which describes the spatial heterogeneity of tissue samples and provide a unique opportunity to establish links between specific tissue types and their molecular profiles.

Exploratory analysis of samples is required when it is not known *a priori* which molecular signals are important for addressing a biological question and so a strategy of detecting as much as possible is undertaken. The process of applying untargeted MSI to explore the chemical environment can be divided broadly into three stages:

1. Image Acquisition
2. Data Analysis
3. Biological Interpretation

Each of the three stages have further subdivisions, as illustrated in Figure 1.1, and as they occur in series the output of a stage can impact on the quality or effectiveness of subsequent stages. This literature review will now detail in turn how each stage of an imaging experiment contributes to the measured data. It will progress in an approximately sequential order, dealing first with sample preparation for image acquisition, second with the of physical process of spectral imaging and finally what data processing is required for information extraction for biological interpretation.

1.3 Image Acquisition

1.3.1 Sample Collection

For useful biological knowledge to be extracted the chemical distributions imaged must resemble as closely as possible the original biological state which makes careful sample collection and preparation critically important. Samples must be excised swiftly so that the natural biological state is preserved and then stabilised so that spatio-molecular changes do not occur[77]. The exact time window for excision and stabilisation depends on the biological activity being considered but is generally between seconds (e.g. ATP degradation[18]) and minutes (e.g. protein autolytic activity[26]). Stabilisation (or fixation) methods have been developed as routine procedures within the histological community and standard practises can be found for most tissue types[44]. Following stabilisation thin sections of tissue are collected and mounted onto a flat target which is then introduced to the mass spectrometer. Not all stabilisation methods have been tested and shown to be compatible with MSI, those that have been are summarised here.

Flash Freezing Freezing fresh tissue blocks by rapid immersion in a fluid cooled well below 0°C preserves the spatial distributions of molecules by encasing them within ice formed from the water content of the tissue. The low temperatures also reduce any enzymatic activity, preventing further molecular degradation[18]. Freezing must be rapid e.g. using cryogenic temperatures, to prevent the formation of large ice crystals which cause cell disruption and subsequently tissue tearing during section collection. Sectioning must be performed on a cryo-microtome cooled below freezing, frozen sections are then thaw mounted onto a sample holder which is at a warmer temperature. Long term storage at -80 °C following snap freezing results in little degradation over time periods of one year but is not advised in excess of this[35]. It has the advantage of being the fastest tissue fixation method and freshly excised tissue can be preserved this way for analysis of both small and large molecules[33]. It can be difficult to obtain sections from tissue that lacks strong internal structure without supporting the tissue block within an embedding medium such as polymer-based compounds(which may negatively impact the quality of the spectra acquired)[202].

Formalin Fixation Formalin fixation is a common histological tissue stabilisation procedure that prevents proteolytic degradation by cross-linking of neutral amino groups using methyl bridges[94]. Fixation is performed by soaking samples in a solution of formaldehyde (typically 4% v/v formalin in water[94]) to which salt buffers are included to achieve an isotonic solution and avoid osmotic damage. This process is performed at room temperature and requires several hours for the fixative to diffuse through tissue[50], fixed tissue

can then be stored at room-temperature[32]. The fixation cross-links and preserves proteins by preventing further enzymatic degradation but due to the time taken to fully penetrate tissue blocks some protein and small molecule degradation can occur[35].

This approach has been shown to be compatible with MALDI MSI of lipids and does not alter the distributions of commonly detected polar lipids[32, 82, 164]. One experimental difference is the predominant formation of sodium adducts due to the sodium salt used in the buffer rather than the protonated or potassiated adducts more commonly detected in tissue imaging [32]. Formalin fixation increases the mechanical integrity of tissue[164] but freezing is required to produce a solid state suitable for sectioning. For protein imaging a process of antigen retrieval is required before protein cans be detected[35].

Formalin Fixation and Paraffin Embedding An additional stage following formalin fixation to improve the longevity of samples is to replace the water contents of the cell with paraffin. The resulting formalin fixed paraffin embedded (FFPE) tissue is then suitable for long term archiving at ambient conditions. The long term use of FFPE has produced substantial banks of historic tissue[227]. Cross-linking between proteins becomes near-complete but most small molecules undergo degradation[189].It has the added advantage that fragile tissue is well supported making it easy to section and sectioning can be performed at room temperature.

During the process, the water content of the tissue is displaced with a series of ethanol washes and infused with paraffin. Unfortunately, many lipids are soluble in organic solvents and this process washes them out of the tissue making this approach suitable for protein/peptide imaging but not lipids[227]. Before imaging can commence, the tissue must be de-paraffinated and antigen retrieval steps undertaken to un-link the proteins[127, 189, 227].

1.3.2 Preparation for MSI

Workflows developed for histology have been successfully adapted for MSI[199]. The use of histological sections for understanding biological function and disease pathology has a long history and is the gold standard for visualising tissue pathology[189]. Adapting these procedures allows comparisons with well understood staining and permits results to be presented in a style that is familiar to bio-scientists[36]. In traditional histology analysis thin tissue sections of $\leq$10 µm thick are collected as this corresponds to the typical diameter of mammalian cells and so captures a single cell layer without optical interference from adjacent layers. This volume is also suitable for MSI but slightly thicker samples can also be imaged (5-100 µm) with minimal variation in ion yield[35, 58, 221]. The sections must be mounted onto a conductive substrate and grounded

to avoid sample charging effects during image acquisition[74]. This can be the standard metal sample holder provided with mass spectrometers or some manufactures produce indium tin oxide coated glass slides. An advantage of glass slides is that optical images can be collected from the same tissue section analysed by MSI[181].

Spatially resolved strategies that do not collect sequential spectra are sometimes used when a whole image is not required or samples are very large. The simplest is just a single spectrum collected from a defined spatial location[30]. Sometimes termed a profile measurement these provide targeted information about a particular region or molecule. Locations known to be characteristic of a tissue type can be identified by biological domain experts[219].

1.3.3 MALDI-Time of Flight (TOF) for Biological Mass Spectrometry

Matrix Assisted Laser Desorption Ionisation (MALDI) mass spectrometry is a 'soft' ionisation technique developed by Karas and Hillenkamp[114] and Tanaka[207] that is capable of producing intact gas phase ions from large (>kDa) molecules. It is the most commonly used ionisation method for Mass Spectromety Imaging (MSI)[2, 74, 223] so discussion on instrumentation will be restricted to this experimental set-up with contrast to other techniques made where appropriate.

Samples to be analysed are co-crystallised with a matrix, as illustrated in Figure 1.2, usually a low molecular weight organic acid, before being interrogated with a pulsed laser. The interaction of the laser with the matrix/tissue mixture causes a portion of the surface material to be ejected, with a small fraction of this becoming gas phase ions. The mechanics of desorption/ionisation as it occurs in the matrix-analyte mixture are still actively under investigation [57, 117, 118] but many factors have confirmed effects on the nature of the spectrum acquired including, matrix choice, matrix deposition and crystallisation, laser wavelength[57], pulse length[57], pulse repetition rate[211], fluence (energy per unit area)[57], spot size (independent of fluence)[74] and instrument parameters. The majority of these are treated as experimental optimisations that must be performed for each sample type[57, 74]. As only molecules that are successfully ionised and transported to the detector can be counted this presents a disconnect between the molecular contents of the sample and the resulting mass spectrum. So the analysis is actually of the cumulative effect of underlying biology; sample collection and preparation; matrix and solvent choice and application; and the MALDI ionisation process which introduces a non-linearity between molecular concentration and the resulting spectral intensity.

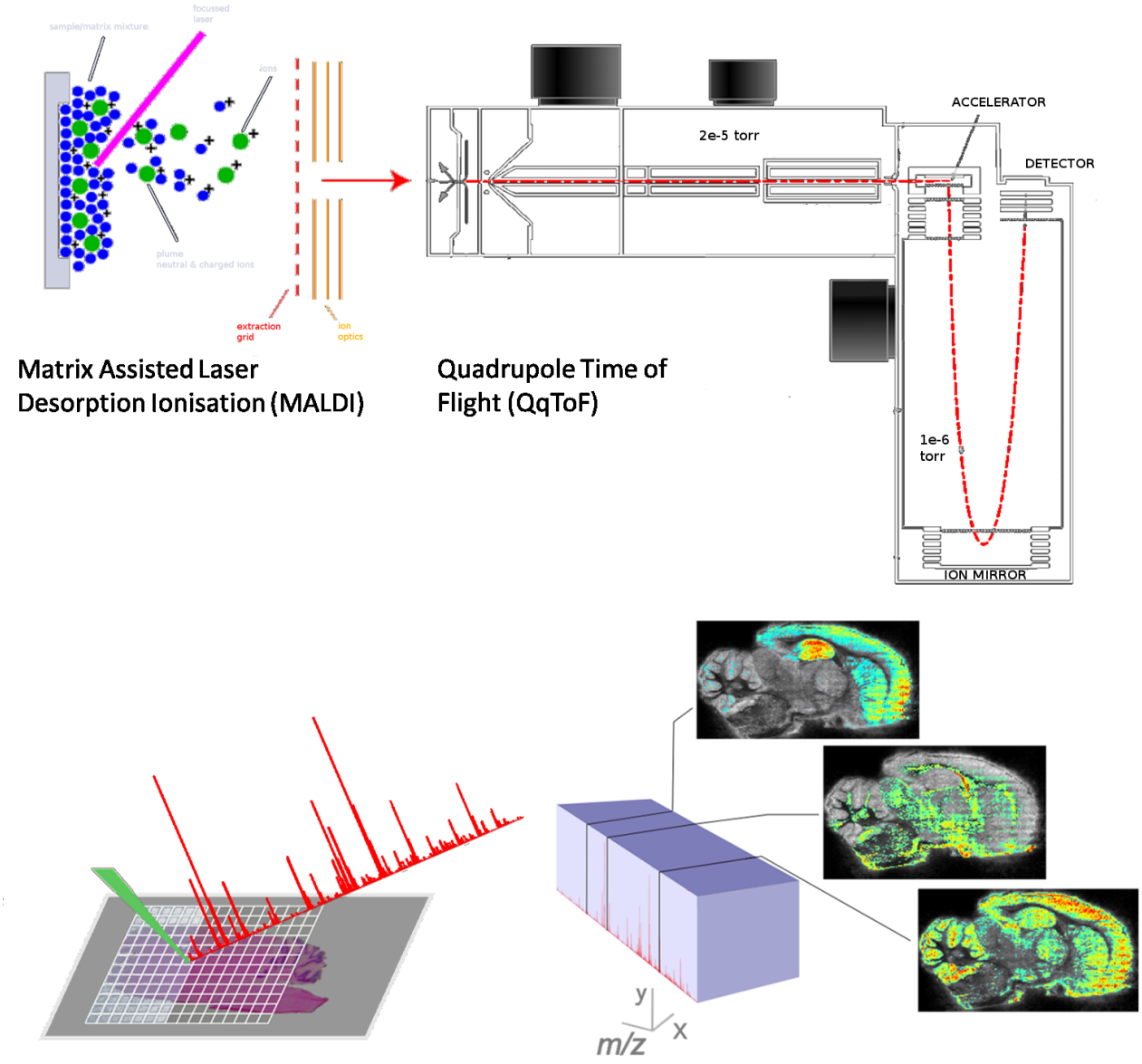

Figure 1.2: Schematic of data collection for MALDI MSI. Top row: collection of a single spectrum is achieved by irradiating a MALDI matrix coated sample with a laser. The resulting ionised molecules are transferred and separated within the mass analyser. An MSI datacube can then be generated by producing spectra over a grid of spatial positions. Instrument diagram modified from [186]

MALDI Matrix Choice and Application

Matrix Choice The choice of 'best' matrix for a particular imaging application is affected by several factors, including, absorption cross-section at the laser wavelength; class of analyte to be detected and spatial resolution. Different matrix and solvent combinations have found applications in the detection of specific classes of molecules[57], for example, α-cyano-4-hydroxycininnamic acid (CHCA) in organic solvent is routinely used for the detection of phospholipids[74] whilst sinapinic acid and formic acid are regularly used in the detection of whole proteins[138]. Crystals of different size and shape are formed (under atmospheric conditions) from each matrix

It is widely acknowledged that for both lipid and protein analysis the choice of solvent system and removal of suppressing species is a determining factor in which molecules are detected[74, 196]. This allows some tuning of the species that are visible in the final spectrum but means that exploratory analysis is not entirely unbiased in terms of the molecules that can be ionised and detected. Choosing a particular matrix affects the final data both in terms of the classes of analyte detected but also the spectral signatures of the matrix itself which may be isobaric with interesting endogenous ions[187]

Matrix Application Analyte detection is also critically affected by the application of the matrix to the sample. Sufficient solvent extraction and mixing must occur for the molecules within the tissue to be successfully ionised but wetting the tissue can cause molecules to be washed out or dislocated and thus confusing the analysis of spatial localisation[53]. Solvent-free mass spectrometry has been used successfully in the detection of phospholipids and proteins which does not risk any dislocation but limits the species accessible[171]. An additional constraint for MSI is that the matrix must be applied homogeneously across the sample with crystal sizes that are smaller than the spatial resolution required[14]. Inhomogeneity in the matrix can cause artefacts such as 'hot spots' are seen in the final data and require suitable compensation[2, 53]. Automated tools have been developed and are available commercially but it is still very common for the matrix to be applied by hand using a nebuliser or an airbrush[32].

Mass Resolution

Mass resolution describes how precisely the mass of a particular ion can be measured and is described by two key measures: mass resolution and mass accuracy. w_{FWHM} is the experimental Full Width Half Maximum (FWHM); m_0 is the accurate m/z of an ion (calculated from its elemental composition) and m_{exp} is an experimental peak centroid m/z

Mass resolution at FWHM, m_{res50}, is defined as

$$m_{res50} = \frac{w_{\text{FWHM}}}{m_0} \tag{1.1}$$

Mass accuracy is defined as

$$\Delta m_{ac} = |m_0 - m_{exp}| \tag{1.2}$$

This is typically quoted relative to the accurate mass in units of parts per million (ppm):

$$\Delta m_{ppm} = 1e^6 \frac{\Delta m_{ac}}{m_0} \tag{1.3}$$

One of the reasons for the near-ubiquity of TOF mass analysers is MSI is that they provide a good compromise between resolution, speed and cost[2, 41] and can be used across an almost unlimited mass range allowing for a flexibility of application. TOF mass spectrometers typically have mass resolutions in the range 10,000-60,000 (the lower end for linear-TOF devices and the upper achieved with QqTOF technology) which allows separation of peaks from isotopes of a single molecule in the low mass range but is not sufficient alone for exact-mass identification[93, 186]. Higher mass-resolution analysers have recently been coupled to MALDI producing images with a mass resolving power of e.g. $> 1,000,000$ [179].

1.3.4 Mass Spectrometry Imaging

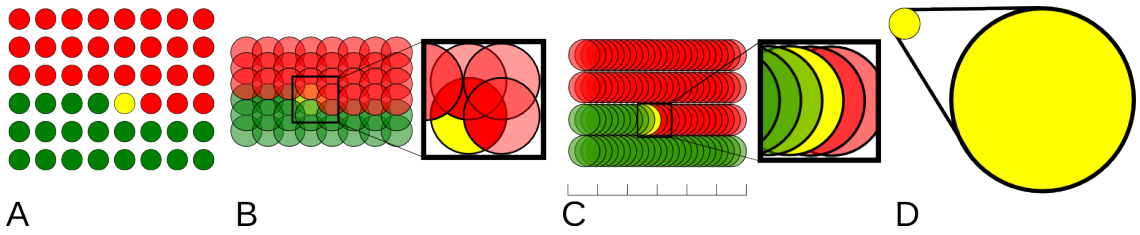

Figure 1.3: Modes of laser sampling for MS image acquisition. Each spot represents a laser interrogation point, red spots have been irradiated, yellow spots are currently being interrogated and green spots are yet to be sampled. A: microprobe image mode - spots are collected sequentially and the spatial resolution is defined by the diameter of the laser beam. B: Microprobe mode with oversampling - the laser spots overlap so only the small area of fresh material is sampled (see zoom box). C: Continuous raster - the laser fires continuously whilst the sample is moved underneath it so it is decoupled from lateral spatial resolution. Pixels are defined by spectrum collection time and this determines the lateral resolution. The laser D: microscope mode - the spatial distribution within a laser spot is magnified using ion optics so the whole image is within a single laser spot.

By confining the area from which the molecules are ionised, and recording the location, imaging spectrometry can be performed, the common modes are illustrated in Figure 1.3. In all cases a focussed laser probe is used to interrogate the surface of the sample producing signal only from areas coated with the MALDI-matrix. In microprobe mode the image acquisition collects a whole spectrum at each location[200], whereas in microscope mode a few spectral channels are collected sequentially with spatial resolution within the laser area[110]. Overlapping the laser pulses so that the matrix is completely ablated in earlier spots and then advancing the laser by a sub-beam radius distance means that signal in subsequent spectra come only from the small area of un-irradiated matrix[111]. This allows the resolution to be increased without requiring hardware changes but can decrease the sensitivity of the measurement as the laser energy is spread over areas that have already been sampled. Continuous raster uses a high repetition laser that continuously fires as the sample is moved underneath. The instrument continuously acquires spectra and pixels are defined by the time during which each spectrum was collected (linked back to spatial position through knowledge of the sample movement rate). This decoupling reduces instrument processing overhead so data can be collected rapidly[196, 211]. Raster lines are typically collected independently. Both ultraviolet (UV)[207] and infrared (IR)[159] lasers have been presented for the desorption of intact biomolecules but UV lasers are more routinely used.

Regardless of the collection mode the dataset resulting from imaging consists of a spatially resolved collection of spectra which can be considered as a 3D data cube, where each entry has two spatial co-ordinates and one spectrum (or 4D if the third spatial co-ordinate is collected), as illustrated in the final panel of Figure 1.2. Temporal artefacts can emerge in the data following sequential acquisition[198] which may require either careful sample stabilisation and/or subsequent normalisation during data processing[53].

Spatial Resolution

Spatial resolution describes the length scale at which MSI data is collected. Increasing the spatial resolution (i.e decreasing the pixel size) allows smaller image features to be seen and so must be chosen with this in mind, for example, the internal organs in small rodents are on the order of mm to cm whilst substructures can be sub mm and individual cells have a diameter of $\approx$ 5-20 µ[111]. Strictly speaking, resolution refers to size of features that can be resolved in an image but in the MALDI MSI literature typically only the pixel dimensions are reported.

Most MALDI imaging spectrometers use microprobe mode, as illustrated in Figure 1.3 where either the beam focus or oversampling is used to control the irradiation area. Decreasing the focus size of the laser

through improved optics has been successfully used to image with pixel widths of around 1 µm[236]. Sub-beam resolution imaging using oversampling with a 200 µm diameter laser is capable of acquiring pixel of width 25 µm[111]. Most commercially available instruments offer laser spot sizes of 10-200 µm[2, 74]. MALDI imaging is approaching the resolution required for the routine generation of cellular scale images.

The largest trade-off with increased resolution is a decrease in sensitivity; as smaller sample areas are interrogated there is less material to contribute to the signal. Increased focussing using beam-stops necessarily removes energy from the laser and oversampling expends beam energy into already ablated areas. A decreased fluence at the surface results in a lower ion yield and, again, reduced sensitivity[117]. Additionally, tram-line artefacts can be produced from the non-uniform beam profile during oversampling[111, 197]. A limit on the spatial-resolution (as opposed to pixel size) is determined by the size of the matrix crystals. As analyte is incorporated within the matrix during application attempting to image at resolutions smaller than the crystal produces matrix crystal-shaped image distortions[20]. One other point to highlight is that collecting data with smaller pixel sizes produces many more spectra per unit area (the number increases with the square of the decrease) so imaging experiments take longer and greater data volumes are generated.

1.3.5 Comparison with Other Chemical Imaging Modalities

Alternate chemical imaging modalities exist which also provide some ability to differentiate a specific chemical distribution within a scene. Figure 1.4 presents an overview of their spatial resolutions and chemical sensitivities. This does not provide an exhaustive comparison of imaging but includes the majority of imaging techniques which are used on histological tissue samples to put the capabilities of MALDI-TOF-MSI into context.

Secondary Ion Mass Spectrometry (SIMS) is the most mature MSI technology[22] which uses a focussed beam of ions to irradiate the sample surface. The ion beam can be focussed to a nm width area and so SIMS can achieve very high spatial resolutions but it is a harder ionisation source so there is significant fragmentation of large molecules. It can be used to map elemental distributions as well as fragments of biologically relevant molecules (such as lipid head-groups) with sub-cellular resolution[160].

Raman microscopy is sensitive to the vibrational modes within samples, the combination of which can provide a specific molecular fingerprint, in complex tissue samples this cannot provide unambiguous molecular identification but classes of molecules (e.g. proteins and lipids) may be separable[235]. As an optical microscopy technique the resolution is limited by diffraction and the optics of the microscope and so the resolution limit is around 1 µm.

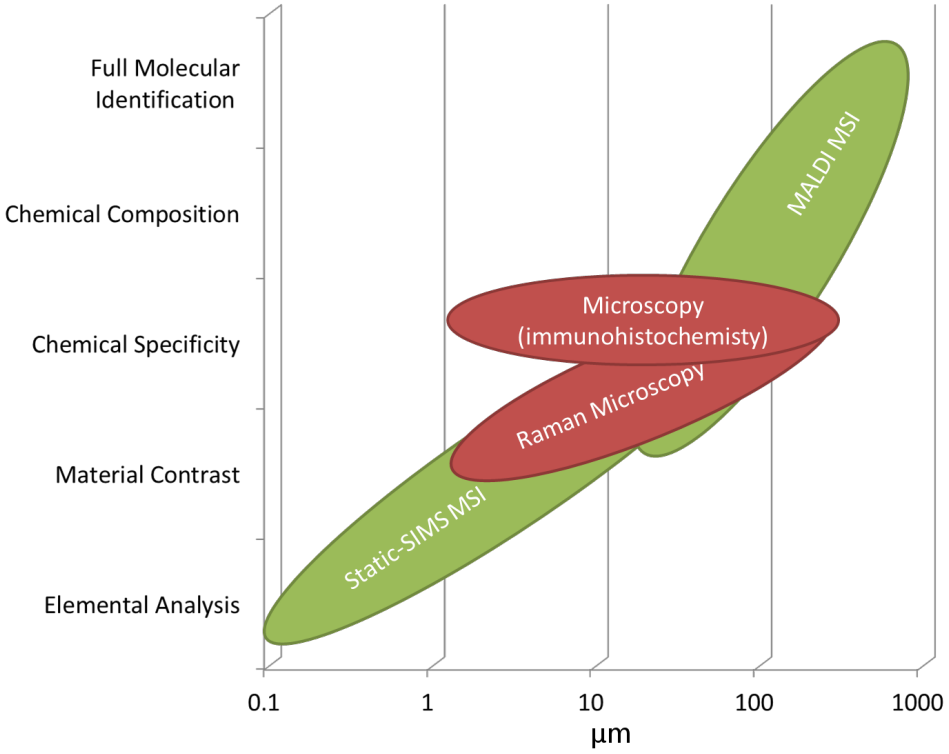

Figure 1.4: An illustration of the spatial resolution (in µm) and chemical sensing abilities of selected commercially available modalities. Green indicates that modalities are able provide multiplexed images with molecule-specific contrast.

Immunohistochemisty uses optical contrast agents attached to labels that adhere to proteins using antibodies raised against specific antigens. The reporter may provide either fluorescent or absorption contrast and, when used in combination with microscopy, sub-cellular spatial resolution can be achieved. Single stains can be viewed with a standard microscope and collecting an optical spectrum can allow the un-mixing of several stains[205] but due to labelling inaccuracies, such as between protein post-translational modifications, unambiguous molecular determination is rarely possible[178]. As the oldest and most studied technique this is the gold standard for understanding disease on a cellular and tissue level and is routinely used for clinical diagnosis(e.g.[155]). However, only a small number of stains can be used concurrently and operator bias is an issue in the interpretation of histological samples[205].

Compared to other modalities MALDI-MSI offers multiplexed imaging of thousands of molecules within the same acquisition and provides chemical specificity over a wide mass range. The current state of technology does not yet provide routine single micro-meter resolution but substantial improvements have been achieved and seem likely to continue to push the boundaries of spatial resolution. For exploratory analysis it has a strong advantage in the number and type of molecular species that can simultaneously imaged without any specific labelling.

1.4 Data Analysis

In a non-targeted MSI experiment one of the principal data analysis goals is to determine trends in the spatial distributions of molecules which can be related back to the underlying biology. An untargeted dataset can contain hundreds to thousands of peaks but it is rare that more than a few molecular masses are identified prior to the imaging experiment[203]. This can lead to a situation where analysis focusses on a small collection of known ions and more complicated inter-molecular patterns are ignored[203]. From a human interpretation perspective, manually identifying trends involving more than a few variables is extremely difficult so computational algorithms are required to extract more complicated trends. The methods must be demonstrated to be robust so users have confidence in the output. Automated feature identification approaches have been developed as routes to minimise manual investigation and to provide standardised protocols for routine analysis[4, 6, 146, 148, 201].

In this section, the data processing pipeline is examined one stage at a time and the core routines that have been developed for extracting spatial and spectral features from MALDI-MSI datasets are discussed. As discussed in Section 1.3.4, MALDI-TOF in microprobe mode is the most common instrument in use for biological imaging studies and so most spectral processing pipelines have been tailored for this data

type[2, 41]. As the data is collected spectrum-by-spectrum the data modelling must account for inter-pixel variability as well as intra-spectrum noise.

1.4.1 Data Description

A standard set of notation for describing spectral data is defined here.

Notation Matrices are denoted in upper case bold; vectors are denoted in lower case bold; and elements are denoted in lower case italic. Matrices, vectors, and elements from the same matrix all use the same letter (e.g., $\mathbf{A}$, $\mathbf{a}$, a). The transpose operation is denoted by the superscript $\top$. The identity matrix is denoted by $\mathbf{I}$. Vector and matrix sizes are denoted by subscript italics.

Data Matrices A whole image is stored as a two-dimensional matrix $\mathbf{X}_{m\times n}$ where it is explicitly assumed that a consistent set of m/z values were acquired for each sample. The jth row vector extracted from $\mathbf{X}$ is denoted $\mathbf{x}_{1\times j}$ and the ith column vector extracted from $\mathbf{X}$ is denoted $\mathbf{x}_{i\times 1}$. Each sample (spectrum) $\mathbf{x}_{1\times m}$ is a row vector with m measurements (i.e. a spectrum); an entire image is composed of n such vectors and the intensities of a single measurement channel (i.e. pixels of an image) are a column vector of size $\mathbf{x}_{n\times 1}$. Every sample has an associated three element spatial vector $\mathbf{s}_n = [s_x, s_y, s_z]$ that is implicitly used when reforming the data into images.

1.4.2 Spectral Pre-Processing

Each of the elements within a spectrum $\mathbf{x}$ contains an numeric intensity value which is the digitisation of a voltage within the detector. The exact values are a combination of the voltage due the number of ions striking the detector and several sources of noise that are independently added to each spectral element. The amount of noise is highly instrument dependent but most spectra require some spectral de-noising (or smoothing) as removing or reducing noise improves subsequent processing[128]. Normalisation is applied to correct for inter-spectrum differences due to systematic fluctuations, e.g. in laser power, detector sensitivity over time, variations in matrix deposition or suppression effects from tissue composition[41, 47, 53, 66, 183].

De-Noising

There are two main types of noise terms that contribute to the detected signal: electronic noise with a Gaussian distribution and 'shot' noise with a Poisson distribution. Electronic noise arises due the thermal motion of electrons within the detector and both the magnitude and variance of the noise are independent

of the detected intensity. Shot noise arises from the statistical process of detecting discrete particle arrivals and produces noise whose magnitude is proportional to the square root of the intensity of the signal. Other possible sources of experimental fluctuations may include transient detector gain variations or statistical fluctuations in ion motion[47].

A study of methods for smoothing (also referred to as de-noising) was presented by *Yang et al*[233] which featured both finite impulse based filters and frequency decomposition approaches. Filtering approaches reinforce the expected peak shapes using a mathematical model whilst frequency based approaches try to identify and remove elements of the signal considered to be noise. The study examined smoothing in combination with peak detection on linear TOF data and it was found that decomposition with wavelets removed high-frequency noise from simulated peptide signals (1-5 kDa) more successfully than impulse based methods but at a cost of substantially longer computation time. When applied to real data all algorithms were found to provide similar sensitivity. Impulse based filters that enhance the expected peak, such as running averages[233], Gaussian and Savitzky-Golay[128, 135], need to be tuned to specific peak profiles but if carefully chosen are fast[233], an important consideration for MSI where the process must be repeated for many thousands of spectra.

Baseline Correction

TOF mass analysers based on analogue-to-digital converters suffer from a baseline artefact caused by the detector becoming overload by large numbers of molecules. The result is a characteristic decaying baseline function superimposed onto the spectrum[59, 142, 226]. Most methods for removal involve estimating its intensity and then subtracting that signal from the spectrum, methods for estimation use either 'dead zones' between peaks[142]; a moving window larger than genuine peaks[226]; or by rejecting these frequencies during wavelet filtering[59]. An advantage of wavelets is that they can simultaneously remove high and low frequency noise so performing de-noising and baseline correction simultaneously. The baseline artefact is absent from mass analysers that include an orthogonal mass separation stage such as the quadrupole-time-of-flight (QqTOF)[93].

Normalisation

The application of normalisation is routine in many areas of mass spectrometry, especially for quantitative analysis where normalisation is made against an internal standard that has identical ionisation and transmission characteristics as an analyte of interest[232]. For untargeted and exploratory analysis there is not

always a practical internal standard and so proxy estimates of global spectral variation have been considered. There is still debate as to the most appropriate normalisation to use but Total Ion Chromatogram (TIC) normalisation has been shown in certain cases to produce images with fewer visual irregularities such as hot spots[53]. However the same study also showed that the use of TIC normalisation could introduce artefacts so care and visual assessment was recommended during its application. As TIC normalisation makes the assumption that the same number of ions will be produced from each spatial location it may be difficult to justify its use when a limited mass range is collected and alternative intra-spectrum metrics could be considered. Within the literature these include normalising to quantiles, $\max(\mathbf{x})$, $\min(\mathbf{x})$, range $(\max(\mathbf{x}) - \min(\mathbf{x}))$ which were all found to improve classification of surface enhanced laser desorption ionisation TOF spectra compared to no normalisation[153].

Spectral Calibration

Systematic shifts that affect the whole spectrum (e.g. due to instrument fluctuations) can be compensated for with re-calibration or on-line lock-mass calibration using identifiable ions within each spectra[149] but this cannot compensate for spectrum-to-spectrum variability in the centroid m/z detected for an individual peak due to statistical fluctuations. Peak modelling on data from high-resolution instruments can reduce the centroid error to single parts-per-million[132] but this approach requires confidence that each peak is produced by a single ion species, an assumption that does not hold for intermediate or low resolution mass analysers.

1.4.3 Spatial De-Noising

Spatial de-noising aims to smooth the variation between spectra that cannot be compensated by normalisation alone. The assumption is made that adjacent pixels will share some spectral similarity due to them containing similar chemical environments. Conceptually, this is a sensible tactic, under the assumption that adjacent pixels have a high likelyhood of consisting of the same region type and therefore producing similar spectra[4].

Many of the software packages for viewing MSI include options for interpolating the pixel values in individual ion images[2]. Edge preserving image de-noising tries to avoid blurring at the edges of distinct image areas and has been applied for MSI, again smoothing ion images independently[4]. Whilst spatial variation no doubt exists within MALDI-MSI it is not established whether it can be estimated on a channel-by-channel basis. An approach for de-noising which takes the whole dataset volume into account has been proposed using a 3D wavelet decomposition[229] but this does not seem to be commonly implemented,

possibly due to the size of MSI data making the application prohibitive.

1.4.4 Dimensionality Reduction and the 'curse of dimensionality'

Following pre-processing the spectra should be largely free of instrument noise and suitable for analysis and interpretation. Data collection on a modern TOF mass spectrometer routinely collects upwards of hundreds-of-thousands of m/z measurements per spectrum. The number of measurements is referred to as the dimensionality of the data. Due to the large number of spectral channels MSI datasets fall into the category of 'high dimensional' or hyperspectral data. However, mass spectra typically contain more measurements (m/zs) than are truly required to describe the data (the lower dimensional data is embedded in a higher dimensional space). As an example in MSI, two adjacent channels within a mass spectrum are often co-linear as they form the slope of a single peak so they describe the same information and add redundancy to the measurements.

The high dimensionality introduces some practical computational restrictions as the size of the image data produced gets very large as the number of pixels increases, mass spectrometry images can regularly exceed 10s of GB per image, but some limitations due to the fundamental mathematical properties of high dimensional data are also encountered. One very practical reason for including dimensionality reduction is to decrease the computation time required to process data of high dimensionality[2], some authors are honest in their pragmatism in selecting a final dimensionality which "provided manageable covariance matrices"[201]. A mathematical aspect that suffers as the dimensionality grows large is that the measurement of distances between samples becomes very imprecise. It will be seen in Section 1.4.5 why accurate distances are important for clustering. Including all m/z measurements in distance calculations does not improve the solution obtained but merely makes analysis of distances between data points more difficult. Collectively, problems related to the number of channels collected are referred to as the 'curse of dimensionality'[70]. Dimensionality reduction methods have been developed for addressing the curse of dimensionality and when used appropriately have been shown to lead to better data visualization and improved classification[38].

Data Reduction vs Dimensionality Reduction

There is a subtle but important difference between the two approaches for addressing the 'curse of dimensionality': data reduction and dimensionality reduction. Data reduction is simply a response to the intractably large data volume and discards some portion in order to make processing the remainder possible. Dimensionality reduction aims to locate a smaller subspace of the data that preserves all of the information by combining

redundant measurements (and possibly discarding measurement dimensions that are purely noise). With MSI data the situation is such that the data is collected in many more channels than differences truly exists within the data: this forces the data onto a higher dimensional space than required so some dimensionality reduction is appropriate.

A summary of the methods used for restricting the data volume for processing is presented here.

Data Reduction

Channel Selection Mass spectra are always collected along a large but finite number of spectral channels as a m/z range is defined for data collection. This is necessary as it is impossible to optimise instrument tuning for an infinite mass range but practically the range will often be further limited to a particular set of analytes of interest. It is possible to range restrict after collection to reduce the computational demands and this is often done with prior knowledge, for example that common analyte classes fall within a certain mass range[32]. Approaches which discard pixels have also been proposed[192], but as the m/z dimensionality is usually greater that the spatial, spectral reduction is more commonly encountered. The only channel selection criteria guaranteed not to discard information is the elimination of zero variance channels which, by definition, contain no discriminating information but some schemes for automatically determining 'informative' channels have also been shown to make on-tissue spatial patterns more pronounced[66]. Selecting a range of channels to collect data over is inherently a data reduction strategy as it decreases the number of channels that are collected at the expense of possibly excluding useful information.

Re-Binning The data can be re-sampled onto fewer measurements by combining mass bins or interpolating onto a lower resolution measurement axis. Re-binning has been performed in both the spectral[67, 87, 192] and spatial[87] domains, with the spectral being much more common as this is the domain of highest dimensionality. In cases where the data measurements are collected at higher frequency than changes in the spectra, as evidenced by peaks spanning tens of m/z channels, it is possible that a certain degree of spectral re-binning can be tolerated with little loss of signal. The main danger of re-binning at fixed m/z intervals is that peaks become merged, reducing the spectral specificity. This is an aspect that is currently under-researched in MSI and will become increasingly important as hardware development continues to increase the m/z measurement density.

Dimensionality Reduction

Dimensionality reduction techniques extract, by some criteria, the salient features from within the dataset and forms a reduced set of feature vectors that capture the informative portions of the signal whilst discarding the noise and redundant measurements. Feature detection approaches look for features (peaks) based on models of their shape then feature extraction is used to combine features that show similar patterns.

Feature Detection (Peak Picking) The features within a spectrum are the peaks caused by ions and feature detection is the process of identifying individual peaks within a dataset. By extracting the features, the dataset size and complexity may be reduced, particularly if noise can be discarded during the process. Some type of feature selection is specified as a precursor to the majority of multivariate schemes for mass spectrometry image data processing described in the literature, usually peak detection[78, 91, 129, 147].

Feature selection by peak picking takes advantage of the highly peaked nature of the data and identifies ions by finding local maxima in a spectrum. Maxima finding is typically either gradient based or uses a sliding filter but the quality is heavily influenced by noise within the data[233]. In a comparison by *Yang et al*[233] it was determined that a model based peak fitting, in combination with wavelet pre-processing, provided the best peak detection on individual spectra but was also a slow method that took more than 10 seconds per spectrum. To make slow algorithms practical, a subset of spectra can be used (e.g. 10%[4] or 15%[210]). Further processing of the peak list can be performed to further reduce the list, such as specifying a consensus frequency where only peaks that are detected in sufficient spectra are kept[4] or by peak intensity (e.g. keeping 200 most intense peaks[175]). Peak detection provides a list of peaks for each spectrum independently which must then be compiled into a dataset list. This process of matching peak centroids is called peak alignment. This presents a data-processing challenge for MSI where the spectra are often complex with overlapping peak patterns, additionally experimental variation means that spectrum-to-spectrum shifts in the m/z location of the maxima of a particular ion peak will be observed[2, 4, 41].

Several summary spectra have been proposed for peak detection to decrease the number of times peak picking is performed and generate a high Signal to Noise Ratio (SNR) spectrum[42, 148]. The mean spectrum calculates the average intensity of each m/z channel across the dataset so averages out both noise and peak centroid fluctuations but suffers from a bias towards intense and common peaks, the basepeak spectrum evaluates the maximum of each m/z channel to maintain intense but infrequent peaks, TIC normalised versions of the mean and basepeak spectra try and reduce the bias towards high intensity peaks[42, 148]. Recent work has attempted to identify peaks by looking at measures of ion image heterogeneity[3]. Once

compiled and filtered, the final m/z centroid list can be passed on to further processing stages (such as channel extraction, discussed in Section 1.4.4).

Deconvolution Deconvolving a spectrum uses a complete model of the ion detection process to computationally reverse the signal generation by minimisation to solve for the parameters of a function describing the spectrum. Gaussian peak shapes are often assumed[4, 212] and a few individual peaks are typically independently identified[2] rather than fitting for all peaks within the data. This approach takes the most sophisticated view of the data but is also the slowest and as such only a few peaks from a subset of spectra can be extracted in a reasonable amount of time[210].

Feature Extraction Feature extraction is a process of concentrating the measurements identified by feature selection into sets of trends within the data, so the number of redundant measurements is reduced. Even following feature detection MSI data can still contain high levels of redundancy, e.g. multiple peaks appear from a single molecule due to the existence of isotopes and adducts, despite experimental protocols for reducing spectral simplification though use of additives[81] or matrix application[185]. Most common approaches for dimensionality reduction are difficult to implement computationally on MSI data and require an initial stage of reduction[172].

Feature extraction fits into the MSI workflow in several important ways: firstly, it can directly expose biological relationships[108] (see Section 1.4.5), secondly, it can improve the response of further computational processing stages by circumventing the curse of dimensionality[166]. Two methods are predominantly encountered within MSI data processing: peak selection and factorisation, the motivation and implementation of each method is discussed here independently but many pipelines found in the literature require both. Factorisation combines spatially covariant molecules whilst peak selection chooses peaks based on their predominence within the data.

Peak Selection Peak selection uses one of the peak detection methods (see Section 1.4.4) to generate a list of peak centroids. The exact peaks chosen will depend on the detection method and which spectra are used as an input. This list can still be longer than is manageable for data mining and contain small or noisily distributed peaks. Frequency filters are applied to keep 'consensus' peaks that appear in a fixed percentage of pixels[4] so this method filters based on magnitude and frequency. It was found that a small subset of the peaks detectable within an image still allowed for effective extraction of spatial information[5] and that the remainder could be queried subsequently for correlation so it may not be necessary to maintain all peaks so

long as informative ones are kept.

An alternative pipeline reduced the peak lists by evaluating spatial correlations against 'non-informative' (usually MALDI-matrix) ion distributions and argued to retain a more 'informative' dataset[66]. If the number of peaks is larger than 100-200[2] then further feature extraction for dimensionality reduction may be required before information visualisation is possible. Once the final peak list is produced a datacube can be produced by generating an ion image for each m/z in the list.

Factorisation The general motivation for performing a matrix factorisation is to express the input data in a particular form that exposes its structure and properties in a clear way. A matrix factorisation, as a mathematical entity, decomposes an input matrix into the product of two or more matrices. Many different factorisations exist with a variety of constraints on the number of product matrices used for the decomposition alongside numerical constraints on individual product matrices, it is the constraints that make these decompositions useful tools. Matrix factorisations are the basis of many classical data analysis methods but often cannot be applied to high-dimensional data due to the computational cost, preventing well-understood techniques from being used to provide insight into these datasets.

Principal Component Analysis (PCA) is probably the most commonly encountered factorisation as it simultaneously reduces dimensionality, reveals patterns in data variance and suppresses noise with a simple data model[87, 91, 116, 146, 147, 192]. Described first in 1901 by Pearson[169] PCA locates a set of orthogonal transforms that reorients the data onto a new set of perpendicular axes in a manner that the axes are aligned along directions of maximum variance within the data[105]. The axes are ordered so that they describe decreasing amounts of variance, so that the initial few axes describe the majority of the data variance. The complete set of principal components, of which there will be an equivalent number to the dimensionality of the data, is known as the Karhunen-Loéve transform but it is more common to discard later components.

The most common method for locating the principal components is through eigenvalue decomposition by calculating the Singular Value Decomposition (SVD) of the mean centred data $\tilde{\mathbf{X}}$ (where mean-centring is achieved by subtracting the row mean, $\mathbf{x}_{m\times 1}$, from every column of $\mathbf{X}$). The SVD of $\tilde{\mathbf{X}}$ gives:

$$\tilde{\mathbf{X}} = \mathbf{U}\Sigma\mathbf{W}^T \tag{1.4}$$

where $\mathbf{U}$ and $\mathbf{V}$ form an orthonormal basis for the rows and columns of $\mathbf{X}$ respectively, the vectors in $\mathbf{V}$

provide the principal component coefficients. Σ is a diagonal matrix describing the square of the variance contained in each component ($Phi^2 = Sigma$).

Once the principal components coefficients (eigenvectors) have been discovered and a variance threshold established, the data can be projected onto the retained coefficients to calculate the scores for each component,

$$\tilde{\mathbf{X}} = \mathbf{VB} \tag{1.5}$$

These scores will tend to reveal trends within the data that are shared by multiple variables and are ordered so that the highest variance eigenvectors appear first in $\mathbf{V}$, once sufficient variance has been captured the remainder of the components are likely to be noise. The subset of $\mathbf{V}$ maintained, containing the first k eigenvectors, is denoted $\mathbf{V}_k$. Typical thresholds are $> 91\%$ of variance preserved[87]. The data model for PCA requires only that the noise present can be described by a normal probability distribution[192] and only takes into account linear relationships between variables[21].

The trend revealing nature of PCA has been widely applied for the exploration of mass spectrometry data[56, 113, 168] and the interpretation of mass spectrometry images[24, 67, 87, 174]. Several authors claim to have difficulty in interpreting the coefficients produced by PCA[52, 87], as they do not provide univariate *m/z* profiles of specific image regions but by nature model more complex spectral interactions. It is routinely used as a dimensionality reduction prior to classification or further visualisation [67, 113]. PCA is very sensitive to scaling[52, 213] but it is not clear if there are general rules for specific spectral types or if this sensitivity is purely sample dependent.

Independent Component Analysis (ICA) maximises the statistical independence of the bases produced, so that no projection provides any information about other projections[40, 192]. It has found applications in blind source separation[192].

The objective is to achieve a factorisation

$$\tilde{\mathbf{X}} = \mathbf{A}_k\mathbf{S}_k \tag{1.6}$$

where the rows of $\mathbf{S}$ are statistically independent, k controls the number of components produced. The *fastICA* algorithm[98] determines $\mathbf{A}$ and $\mathbf{S}$ from the uncorrelated projections produced by PCA, $\mathbf{B}$. First, 'whiten' the projections so they each have unit variance $\mathbf{Z} = \Phi^{-1}\mathbf{B}$ and then perform an optimisation to

obtain a rotation matrix $\mathbf{U}$ that maximises the kurtosis of $\mathbf{U}^T\mathbf{Z}$, to give

$$\mathbf{S} = \mathbf{U}^T\mathbf{Z} = \mathbf{U}^T\Psi^{-1}\mathbf{B} \tag{1.7}$$

Substituting into Equation 1.5 gives

$$\tilde{\mathbf{X}} = \mathbf{V}_k\Phi_k\mathbf{U}_k\mathbf{S}_k \tag{1.8}$$

Choosing k presents a problem as it is rarely known *a priori* and the projections come out unordered so manual inspection of all factors is required but despite these shortcomings it has been successfully used for extracting protein patterns from mass spectra[87, 141].

Non-negative matrix factorisation (NNMF) is a matrix factorisation approach that performs the decomposition of the (non-negative) data matrix $\mathbf{X}_{m\times n}$ into two other non-negative matrices $\mathbf{A}_{m\times r}$ and $\mathbf{Y}_{r\times n}$ so $\mathbf{X} = \mathbf{AY}$. It is a factor analysis technique which means that a model is required, to which the data is fitted. This model necessarily includes the number of factors present (a value for k, $k < \min m, n$) and, in this case, the non-negativity constraints. The factorisation is solved by minimising the residual[131]:

$$\min_{E,G} ||\mathbf{X} - \mathbf{EG}||^2 \qquad \text{s.t.}\mathbf{E}, \mathbf{G} \geq 0 \tag{1.9}$$

The resulting matrices are referred to as the basis matrix, $\mathbf{A}$ and the abundance matrix $\mathbf{Y}$ so that every data point is the linear combination of the basis vectors multiplied by their respective abundances. Solving the NNMF minimisation to find the best global solution can be difficult as it is very sensitive to starting conditions and so can become 'trapped' in locally optimal solutions[131]. Probabilistic latent semantic analysis has also been demonstrated for pattern extraction from MSI[87] and has been shown to be equivalent to NNMF with a Kullback-Leibler divergence measure[72] so also decomposes the data into two positive values matrices within which the probability of a pixel belonging to a factor is contained. It has been suggested that identifying single ion patterns is easier from positive coefficients[87, 175] and that positive scores present a more human-friendly tissue annotation (compared to PCA & ICA which, being covariance based, can produce positive and negative values). NNMF has been applied for analysis of tumour images in 2D[87, 108] and in 3D tissue[175] but all datsets required substantial reduction through peak-picking before analysis could be performed.

All of these data and dimensionality reduction techniques return a reduced set of measurements that are more mathematically approachable and computationally manageable. The reduced measurement set are then passed on for information extraction and knowledge generation.

1.4.5 Information Extraction

The aim of information extraction is to present the user with a collection of images and/or spectra which show important features within the data. Both the style of final output and how importance is measured can depend on the final application. In the field of biomedical imaging the following types of experiment are the most common.

- Mapping of pre-determined molecules such as the distribution of a dosed drug[28]
- Identification of regions of spectral similarity for exploratory analysis[4, 108]
- Disease/normal comparison (biomarker detection)[19].
- Association of molecules with particular tissue regions or types (profiling -omics style systems study)[123].
- Determination of individual cell types (sample classification)[150]

The endpoint of all of these examples are an image (or a small set of images) with a list of associated m/z values. Within the context of a tissue imaging experiment spectral patterns are usually interpreted as 'biological features' but, as discussed in Section 1.3.4, the tissue biology is only one contributing factor to the ion packet produced and the subsequent molecular signal detected. Any spectral patterns identified are therefore patterns within the ion packets generated, which includes the underlying bio-chemistry; desorption and ionisation processes; and detection and measurement processes of the mass spectrometry.

Ion Images

The first data visualisation method developed for MSI was to map the intensity of the area under a chosen peak to produce an ion image[31]. Such an ion image is defined by the peak centroid (chosen manually or algorithmically) along with a window width. This m/z range is summed to an intensity value in each spectrum and an image formed by rearranging the intensities onto the spatial coordinates. This approach relies on only a single ion contributing to the channels selected and disregards multiple signals from the same molecule (e.g. isotopes, fragments, different charge states). The ion images are usually shown with a false colour-scale linking colour to intensity. The images often show high heterogeneity due to the noise from uneven matrix and laser response causing 'hot spots'. Spectrum-by-spectrum normalisation (see Section 1.4.2) can reduce these effects but often this is insufficient and image post-processing is required[2]. This usually takes the form of limiting the intensity scale to some user-defined maximum and minimum so that pixels whose intensities fall

out of this range are given the maximum or minimum colour respectively. Often this is performed manually but can use a more principled approach (e.g. setting the top 5% to the maximum[223]).

Linear Mixture Models

Mixture models treat the data as if every spectrum can be considered the combination of a set of pure reference spectra[212]. This is analogous to the problem of blind signal un-mixing where it is assumed that a dataset is constructed from a mixture of a small number of unknown 'pure' spectral profiles in unknown proportions, the goal is to elucidate the profiles and the proportions. It is hoped that these unmixed spectral profiles then represent patterns of bio-molecular significance[107].

Concentrating measurements using dimensionality reduction analysis, described in Section 1.4.4, is one popular way of approaching the linear un-mixing problem and has been widely used for interpreting MSI data [2, 52, 108, 166, 192]. Factorisation inherently decomposes a dataset into a linear mixture (as the linear product of spectral and spatial matrices). However, there is currently no consensus on what the best factorisation to use with MSI[2], or even what metrics of success might be appropriate.

Factor Analysis Some prior knowledge is required for applying factor analysis, including a model which incorporates physical constraints on possible measurements and the number of factors. Non-negativity of both spectral and spatial mixtures has been proposed as a physical constraint on the reasoning that only positive valued spectra are detected[87, 192]. Automatic methods for determining an optimum number of components using the Akaike information criteria have also been presented[87]. Despite these developments there has not been a universal uptake of these methods, possibly due to the difficulty of choice of parameters for factor analysis but also due to the difficulty in finding good solutions computationally to the equations detailed in Section 1.4.4. Most algorithms start with a pseudo-random initialisation which can strongly effect the final set of factors[72]. This can require that the algorithm be run multiple times and the best solution kept, which multiplies the computational time. It is also possible that multiple mathematically optimal solutions exist for a given factorization problem[231], which may further confuse biological interpretation). Factorisation using ICA and NNMF was found to confine the baseline noise to a single component whereas in PCA it was preserved in each eigenvector[192].

Component Analysis Component analyses, such as PCA, operate on purely geometric constrains, rather than model based, which can make them more robust in operation, however, prior physical knowledge cannot be explicitly modelled. As the components are determined by minimising the co-variance (PCA) or mutual

information (ICA) between the components they can be formulated so there is a single unique solution[137], removing some ambiguity. Any interpretation is based purely on the hope that biologically significant differences within the image will produce the most variation in the signal which will then be reflected within the components. PCA and ICA both produce negative values for the coefficients, which correspond to negative covariance, and are considered to be 'non-physical' by several authors under the aforementioned assumption that ion formation is a purely additive phenomenon. Non-negative formulations of PCA have been developed[162] but so far have not been evaluated for dimensionality reduction or interpretation of MSI. It has been commented within comparative studies that PCA produces more 'noisy' projections compared to factor analysis, this is probably due to some noise being equally distributed over all the components rather than confined to a single factor[192].

Region of Interest (ROI) or segmentation

ROIs divide the image scene into distinct compartments which are believed to be internally homogeneous. The shared properties of spectra within regions can then be compared for molecular differences[146]. The information extraction challenge is in determining where region boundaries are. The term segmentation is used when every pixel within an image is allocated to a ROI (although there is no constraint on using all ROIs in further analysis). By allocating each region a unique colour and then colouring each pixel according to its region membership, a single overview map can show much of the information content of the dataset[4]. The ability to summarise the whole dataset in a single image is a strong advantage of segmentation over mixture models which still produce several maps that must be interpreted.

Supervised segmentation (ROI selection) seeks to characterise the molecules present in an externally identified tissue type, e.g. m/zs specific to that ROI[25] linking function similarity of areas[4] or tracing the outline of volume features in 3D stacks[46].

For proper statistical treatment for evaluating differences between tissue types (e.g. for biomarker discovery) all spectra from a particular type should be combined[107], combining joint molecular information, which may be a more powerful tissue indicator than univariate m/z markers[4, 25, 107].

Manual Segmentation Manual delineation of ROIs requires a biological expert to divide the image into structural units based on their knowledge of the tissue[229]. These regions can be determined either directly on the image space by examining individual ion images or by a simple comparison of ion images by plotting their intensities against each other. An approach that is becoming more popular is histology driven ROI analyses, drawn with reference to a co-registered image of a tissue section stained with a contrast agent (as

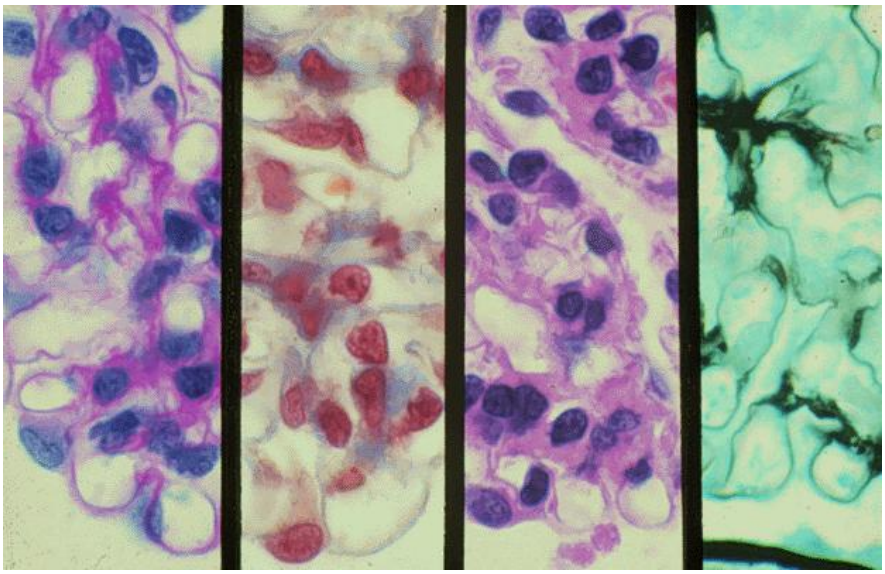

Figure 1.5: A detailed understanding of tissue pathology can require a panel of stains, these renal images show, from left to right: Periodic Acid-Schiff, Trichrome, H&E, Jones silver stains each of which highlight a different type of attribute of the tissue[103] (silver stains collagenous structures such as the glomerular basement membrane; PAS can also accentuate matrix and basement membrane constituents; trichrome can stain immune deposits red and matrix in blue or green, H&E highlights the general structure of cells[144]).

shown in Figure 1.5)[34]. Whilst still the gold standard for labelling tissue images it is labour intensive and any human dependent method has the potential to suffer from user bias and operator variability[25].

Clustering for Automated Segmentation of MSI There has been much interest in developing automated algorithms for segmentation to eliminate the aforementioned limitations of manual selection[5, 25, 54, 67, 146, 221].

Clustering detects sets of spectra that have similar spectral characteristics and groups them together, thus achieving automatic segmentation. This operates on the assumption that spectra collected from a certain tissue type will display similar spectral characteristics that are different from other regions. The challenges for automatic segmentation are in measuring similarity between spectra and then finding optimal groupings. Common similarity measures will be discussed first, followed by algorithms for performing clustering.

Similarity Measures The mathematical 'distance' between two vectors is a formal way of calculating how far apart they are from each other and other vectors within the dataset. This allows a clustering algorithm to identify clusters of vectors that are close to each other and separated from other clusters. The measure is typically chosen so that it is 'short' for vectors with similar patterns of peaks and 'long' for very different patterns. A closely related concept is similarity, where greater values indicate higher similarity. Typically

distance measures can be transformed to similarity measures by subtracting their value from 1.

The most commonly used distance metrics for MSI are the Euclidean (or l_2) norm, vector angle and correlation [52].

The l_2 norm is defined as

$$||\mathbf{a}-\mathbf{b}||_2 = \left(\sum(\mathbf{a}-\mathbf{b})^2\right)^{\frac{1}{2}} \tag{1.10}$$

where $\mathbf{a}$ and $\mathbf{b}$ are two vectors (e.g. spectra). Another vector metric is the angle between the vectors, defined for $\mathbf{a}$ and $\mathbf{b}$ as:

$$||\mathbf{a}-\mathbf{b}||_\theta = \cos^{-1}\left(\frac{\mathbf{a}\cdot\mathbf{b}}{||\mathbf{a}||_2||\mathbf{b}||_2}\right) \tag{1.11}$$

where $\cdot$ is the scalar product of the vectors and correlation. Conceptually, this measures how much the two vectors are pointing 'in the same direction'. Both of these measures return zero when calculated for a particular vector with itself and increase with larger differences. The Euclidean distance is magnitude dependent and so has no upper bound whereas the vectors are divided by their magnitudes in the cosine distance and so it has the range 0-360 degrees. Mathematically related to the cosine angle is the correlation between two vectors:

$$||\mathbf{a}-\mathbf{b}||_r = 1 - \frac{\sum(\mathbf{a}-\bar{\mathbf{a}})(\mathbf{b}-\bar{\mathbf{b}})}{||\mathbf{a}-\bar{\mathbf{a}}||_2||\mathbf{b}-\bar{\mathbf{b}}||_2} \tag{1.12}$$

where $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$ are the average values of $\mathbf{a}$ and $\mathbf{b}$ respectively. Correlation takes the range 1 to -1, with 1 being identical, 0 being un-correlated and -1 being anti-correlated. To turn this into a distance metric the correlation is subtracted from 1. All of these turn two vectors into a single numeric metric which is interpreted as the distance between the points in the vector space.

The Gaussian similarity measure s takes an exponential with the negative l_2 norm

$$s = 1 - e^{\left(\frac{-||\mathbf{x}_i\mathbf{x}_j||_2^2}{2\sigma^2}\right)} \tag{1.13}$$

for a pair of spectra x_i, x_j, σ is a parameter that controls the neighbourhood size. This produces a non-linear decrease in the similarity measure as the distance between the vectors increases.

All of these measures degrade when the dimensionality in which they are calculated exceeds the true dimensionality of the problem[38]. This is where mathematical aspects of the 'curse of dimensionality' begin to effect the processing of MSI data. Under very high dimensionality all distances tend to equal one another and the discriminatory power of these distance measures fails[99].

Parameter Space Various parameters can be chosen to form the vectors that distances are calculated from and the changing the parameter space can change the final clustering result[121]. The key aspects of the parameter space chosen are that it must preserve the differences between datapoints that are important for defining the regions whilst allowing distances to be accurately calculated.

External Information from external sources can be appropriately co-registered with the MS image and then used to generate labels for the MS pixels. Clustering is performed on the external data and region labels can be propagated directly to the pixels. Sources of external information that have been applied for MSI segmentation include histological photo-micrographs[45, 46] and atlases[206]. Advantages of this approach are that these domain usually have few parameters (e.g. three colour channels) so distance calculation is fast and the parameters are independent of the spectral data, so further statistical analysis will not be biased.

It is difficult to ensure that the registration is appropriately accurate. Problems such as image distortion due to the physical geometry of laser delivery during MSI data acquisition[46] or section distortion during sectioning[209]. Even with accurate general registration there may be differences in the resolutions of MSI and histology (which applies to atlases as they are generated from histology[126]) so at tissue type boundaries there will be label competition which must be resolved if a unique label per pixel is required. Clustering schemes which allow partial membership exist but then the aim of simplifying the data into discrete areas is not achieved.

Spectral Calculating the distances directly from the m/z intensity values in the spectra considers all of the peaks within the data. However, as each spectrum can contain hundreds-of-thousands of m/z channels this presents a substantial computational task and so is not typically performed[210]. Clustering on specific m/z grayscale images that reflect most difference between regions could provide a suitable dimensionality reduction and has been shown for SIMS[229]. However, this required manual selection of an appropriate m/z channel and cannot detect regions defined by multivariate changes. A subset of m/z s can be selected (e.g. using peak picking for data reduction)[148] which makes the computation practical and allows multivariate patterns to be detected.

Data transformation Even following data reduction computational time may still be long as many thousands of peaks can be extracted from a given dataset, additionally the dimensionality of the data may remain higher than the true subspace[4]. Transforming from the m/z domain using a component or mixture model (see Section 1.4.4) provides a principled route to reducing the absolute number of dimensions that need

to be calculated upon. As these methods determine a transform that preserves the spectral differences within the data distance metrics can be calculated on the transformed data, with equal accuracy. The computational overhead is transferred to the factorisation task. Of the methods discussed PCA is the most commonly used prior to clustering[25, 212].

Clustering algorithms Much work has been done in developing algorithms for the detection of clusters within data, see *Kriegel et al*[121], for a recent review. The basic problem is to partition the data so that each datapoint (pixel) belongs to a single cluster based on detecting 'lumpiness' in the chosen parameter space[212]. Several algorithms have been applied to MSI and are described here.

k-means The k-means clustering algorithm divides the data into a pre-specified 'k' clusters by minimising the total distance between datapoints and their nearest centroid. It is commonly implemented by iteratively moving the position of the centre (mean) of each cluster of datapoints so that the total distance of datapoints to centres is minimised[101]. Some of the reasons for the enduring popularity of the k-means algorithm are that it simultaneously determined from the clusters and classifies all points whilst being guaranteed to return a result[86]. As it is randomly initialised the output can be sensitive to the start position and so the best of several repeats may be required[10]. Whilst is has the potential to produce arbitrarily bad results, in practise it is found to be generally quite robust[10, 101]. It can be slow as the distance of each datapoint to every centroid is evaluated on each iteration. This algorithm tends to produce clusters that are approximately the same size and of equivalent density[156].

The key disadvantage of using this algorithm is that the number of clusters must be provided beforehand, which may not be known in exploratory data analysis. Expectation maximisation is a similar clustering algorithm that partitions membership based on mixtures of gaussian density pattens[4]. It too requires the number of clusters as an input but allows different sizes and shapes of clusters.

Hierarchical Clustering This algorithm generates a 'tree' of data points which are arranged according to similarity, the terminology follows the tree analogy with the root node containing all data points which then forks out to branch nodes containing subsets of similar points until a set of leaf nodes with a single data point is reached[156]. This tree is built from the bottom up so the pairwise similarity between all datapoints must be calculated[54]. For the number of pixels considered in MSI this can get very large and computational restrictions may be encountered, a top-down pseudo-hierarchical method using repeated application of k-means has been proposed to counter this[210] however this often leads to non-optimal partitioning of the

data[4].

The network is then represented on a dendrogram with the length of the branches representing the linkage between nodes, which is a function of the distances between the data points contained within each node. To form clusters the tree is 'cut' at a particular branch and all the nodes beneath a cut are grouped into clusters. [25]. Clusters can be formed either to a specific number of input clusters, which determines the number of cuts to make or can be approximated from within the tree either by specifying a certain linkage distance requirement to separate clusters or an approximate number of nodes in each cluster. It is common practise to divide the tree by manually inspecting the tree and choosing where to cut[52]. This method provides an optimum separation of the data but the large memory requirements prevent it being applied in practise.

Self Organising Maps The self organising map (Self Organising Map (SOM)), or Kohonen network [119], is a type of structurally linked neural network of that is trained to resemble the distribution of the input data points. Many geometries are available for the network but in this case the map is composed of an equispaced 2D grid of *nodes*, allowing spatial relationships to be described in *node-space*. Each node has a set of weights which correspond to the measurements of the data i.e. the projections onto the approximate basis. Producing a segmentation using a SOM is a two stage process, first the map must be trained to reflect the distribution of measurements found in the data then secondly, the data points must distributed over the map assigning each data-point to the node that is most similar to it.

Spectral Clustering Spectral clustering uses connectivity to collect data-points into clusters, as opposed to compactness measures used by algorithms such as k-means. The tutorial by von Luxburg [220] describes the state-of-the-art of spectral clustering from which the key steps for performing image segmentation are summarised here.

The key idea of this approach is to construct the graph Laplacian for a dataset and then partition it so that only close data points that are left connected. Several variants of 'the' graph Laplacian exist, but the normalised 'random walk' graph Laplacian: $\mathbf{L} = \mathbf{L}_{rw} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{S}$ is recommended by von Luxburg[220]. $\mathbf{S}$ is a similarity matrix containing the pairwise Gaussian similarity (Equation 1.13) for all spectra in a dataset. Multiple measures of similarity exist and can be used but the Gaussian distance is commonly used as the exponent term means that similarity rapidly tends to zero for un-like points so the resulting matrix tends to becomes sparse. $\mathbf{D}$, is the degree matrix, a diagonal matrix where each element of the diagonal contains the total connectivity of each spectrum $d_{ii} = \sum_{j=1}^{n} s_{ij}$. $\mathbf{I}$ is the identity matrix.

The objective of dividing the data based on connectivity rather than concentration is to allow non-compact

shapes to exist, instead of trying to partition the data space into approximately equal sized portions. This may allow small but well separated clusters to be detected[220].

Other data visualisation Recent work has investigated the use of techniques from the field of hyperspectral visualisation to provide an overview of the composition of MSI data[67]. These techniques use non-linear mappings of high-dimensional data onto a three-dimensional space which is then colour-coded with an RGB scheme. The approaches tested included SOMs[67, 119, 228] and t-Distributed Stochastic Neighbour Embedding[67, 216]. As with almost all of the work introducing clustering and visualisation results the results were evaluated against some expected patterns of biological delineation within the tissue sample and the methods produced similar but not identical results. Further work is still required to understand exactly why particular differences are highlighted and how they should interpreted with regards to the underlying bio-chemistry.

1.4.6 Practical Implementation

After a collection of data processing steps has been decided upon it must be put together into a pipeline for the data to progress through. Having a well structured pipeline should make reporting data processing clearer but care should be taken to understand the changes that occur to the data at each stage. These pipelines and their constituent algorithms need to be efficient in terms of both computing and memory requirements to cope with the size of MSI data[2, 172]. Implementations of algorithms that process the data without having to load the complete dataset into memory are known as 'memory efficient'[172]. Imaging datasets are usually stored as sets of spectra so spectrum-by-spectrum processing is typically more efficient to implement but as alternative storage approaches are introduced[182] image processing may become competitive.

Current MSI datasets are large in both the number of m/z and the number of pixels but emerging technologies like high-mass-resolution instruments and 3D imaging are dramatically increasing the size of datasets collected. A general rule has been proposed that the total processing time shouldn't substantially exceed data collection time to avoid this being a bottleneck[2].

1.5 Biological Knowledge

The types of knowledge that can be generated from MSI experiments provide either the spectra (molecular profiles) or specific molecules (e.g. biomarkers of disease state) that differentiate between tissue types. The up- or down-regulation of specific molecules can also be determined if quantitative imaging method are

used[27]. In exploratory analysis the general aim is to simultaneously determine distinct regions and their characteristic molecular profiles.

1.5.1 Molecular Profiles

Molecular profiles provide a signature spectrum from a particular spatially localised region. Historically, this region was defined by pipetting volumes of matrix with enough solvent to wet the tissue directly onto a selected tissue position. The solvent wetting effectively extracts molecules from the area which can then be subjected to MALDI-MS[30]. Profiles can also be produced from imaging mass spectrometry by averaging the pixels selected by segmentation or the factors produced from factor analysis[4]. Comparing molecular profiles has been used to distinguish breast cancer subtypes and normal tissue [113]. It is not necessary to identify molecules in order to plot their spatial distribution, and unknown protein and lipids can be mapped to identify tissue types[46, 210]. Profiles provide a broad compositional view of the tissue which can be further analysed to identify individual peaks within the spectrum that could provide detailed bio-chemical information.

1.5.2 Molecular Identification

In some cases the distribution of individual molecules may distinguish one image region from another. If the identity of these molecules can be definitely established then they may be biomarkers for a disease state[52]. In many cases molecular identification can be established using mass spectrometry. On high resolution instruments this may be achieved using mass alone otherwise further studies of fragmentation patterns[82] or protein digests[36] may be required. In a biomarker discovery workflow, one of the more powerful applications of MSI, the endpoint is typically a list of candidate molecules that show strong covariance with the biological trend being studied. When the intensity of the identified ion is mapped the ion image can then truly be called a molecular image.

1.6 Measures of Success

So far, a collection of tools have been described for denoising, artefact removal and image information visualisation which can be assembled into defined processing pipelines for automated data analysis[4, 148, 210]. To be able assemble a robust and reliable pipeline it is necessary to evaluate a measure of success so that parameters can be optimised. Commonly encountered measures for judging the success of data processing

pipelines for MSI are described here but it is important to note that the definition of success is dependent on the type of question being asked (e.g. hypothesis driven, exploratory or quantitative questions), the measures described here continue to focus on the exploratory questions.

A comprehensive ground truth provides a realistic set of samples that have been comprehensibility annotated[143], where realistic in this context implies that the size, noise distributions and types of annotation are all representative of data as encountered in a natural setting. There is a lack of publicly available annotated MALDI-MSI data so proxy measures are frequently substituted for a genuine ground truth so these are also discussed here. As data-processing becomes more automated it will become very important for the community to have access to shared data as the success of an algorithm can depend as much on the experience of the user as on the nature of the data [151]. The substantial differences outputs from the data pre-processing pipeline require an independent test dataset for comparative identification. Whilst *reproducible* data collection schemes have been reported[8, 161, 210] the relationship between tissue composition and ion signal is not fundamentally well understood. This is exacerbated by the lack of understanding of the effect of the sample handling processes deployed by different laboratories[76]. Consequently a variety of alternative criteria have been used to validate algorithms presented in the literature.

1.6.1 Comparison with annotated histology

Using microscopic inspection tissue sections can be annotated with many attributes such as cell type and disease states and a great number of histological stains exist which make visible architectural or functional cell properties[152]. The use of annotated histology for producing segmentation maps has already been noted but they are also used for evaluating the results of automated segmentation, on the assumption that the most significant molecular differences will correspond to changes in cell type or tissue environment[146].

Histology is the gold standard for tissue research and diagnosis, however, biological samples are highly heterogeneous and disagreements between trained histologists is common[140, 145]. This human variation may be compounded in the labelling of large tissue areas by the disparity in scale between microscopy and whole section imaging[221]. As only small numbers of stains can be used simultaneously[205] there is a limit to the total spatial information that can be obtained from a single section so non-specific stains (often H&E[33]) are used as a partial solution as they gives a wider range to the staining contrast. Using histological staining will only reveal tissue features whose contrast is enhanced by that stain, which can leave other features invisible. Despite these limitations comparison with anatomy visualised by stains is the most common method of determining a 'good' segmentation and the evaluation is made by qualitative

comparison[25, 54, 109, 146]. Distributions have also been compared against a tissue atlas[46, 125, 206]. This poses an issue if the visualisation produces results that are hard to see in histology[108] or samples are analysed that are less amenable to staining[222].

1.6.2 Visual Inspection

Visible structure is a commonly used criteria for discussing image processing results[4, 52, 146] where the end user determines whether tissue structure is reflected in the final output. It relies heavily on the intuition of biological domain experts to make the decision as to whether there is likely to be detectable molecular differences between tissue regions. This is perhaps undesirable as it requires manual evaluation, which makes it unsuitable for processing large numbers of images, and may be susceptible to operator bias. By itself, this is unlikely to provide a valuable source of information even for biological domain experts who do not typically posses an intuition for interpreting mass spectrometry images.

1.6.3 Comparison with Simulation

Simulated data is attractive for evaluating unsupervised methods as it can be completely defined by the researcher[170]. Certain portions of the mass spectrometry process have been well modelled, such as ion transmission though electrostatic potentials[43, 112], whilst modelling of other areas, such as the MALDI ionisation process, is under development[117]. Simulation of single spectra using the model presented in [43] has been used for the evaluation of peak detection algorithms[233]. This model did not include many of the sources of experimental noise present in real-world data[43], which was not a problem for investigating intra-spectra properties (such as peak identification). The absence of inter-spectrum variance is an over-simplification when imaging applications are considered and spectra are being compared as a cohort. Researchers have added Poisson noise to reference spectra to produce a dataset with spectral noise statistics that are representative of real-world ion detection[88, 214]. Other sources of noise, such as chemical noise, statistical sample variation and matrix variability were not included in these simulations.

1.6.4 Data-Dependent Measures

Several approaches for understanding data visualisation algorithms have been used. Rather than evaluation against a ground truth they use side-by-side comparisons of algorithm output. Qualitative assessment of the spatial features visible has been performed and differences observed[67, 108]. The informal use of a cohort consensus was suggested as a method for resolving differences between factorisations[108].

This is similar to the quality metric of segmentation which uses the amount of variation within spectra from a ROI. When the variation is low then a good segmentation has been achieved but this must be offset against a tendency to produce a large number of ROIs with small numbers of highly-similar spectra[25].

1.7 Conclusion

It is the area of exploratory data investigation that has the keenest requirement for advances in computational processing. Major limitations that have been identified are the volume of MSI data, choosing the right algorithms for unsupervised analysis but also presenting the algorithm outputs in a useful style. There is huge potential for molecular imaging using MALDI-MSI and as the tools are developed it will be possible to acquire larger and more sophisticated datasets but also to re-analyse existing data.

A great deal of work has been performed in establishing protocols for the routine collection of MSI data and the acquisition of high spatial resolution images is now quite routine. Computational handling of MSI data remains an obstacle in the routine use of exploratory tissue imaging due to the large number of spectra collected in an image. Determining meaningful patterns within mass spectrometry data is a challenging task that requires the development of specific tools and the evaluation of these approaches. An approach is required that is both computationally and memory efficient so that it can be applied to the large datasets produced by MSI.

Chapter 2

Experiments with Random Projections on Spectral Images

The application of random projections are considered as a solution to the issue of the high dimensionality of mass spectrometry imaging data. Random projections were found to provide a computationally attractive dimensionality reduction which preserved distances between data points such that clustering for image segmentation could be performed.

However, absolute spectral recovery could not be achieved so randomized basis approximation was considered and this concept was developed into a tool for low-loss compression of hyperspectral images. By using this framework, which separates the spectral and spatial variation, a substantial dimensionality reduction was achieved. Further investigation considers effective implementation strategies and a method for combining datasets compressed separately is developed.

2.1 Introduction

The high dimensionality of mass spectrometry images is a major obstacle in data analysis. Too many channels are collected within a single experiment for it be practical for a user to manually inspect all the possible

ion images so automatic processing is required. Many algorithms have been devised for extracting patterns from mass spectrometry data[4, 108, 146], however, the datasets are becoming too large to load at once for computational processing so dimensionality reduction is required.

2.2 Dimensionality Reduction

The requirement for dimensionality reduction prior to further analysis is well established within both the mass spectrometry and the machine learning communities[225] (see Section 1.4.4). Approaches currently used for the interpretation of mass spectrometry imaging tend to consider the problem as a signal processing issue, where an underlying noisy signal can be extracted by direct manipulation of the input spectrum. These approaches reduce the dimensionality by reducing the number of samplings required to describe the signal, either by changing the spectral sampling frequency in the case of rebinning or using a non-periodic sampling in the case of peak extraction.

An alternative view of the data is that of a lower dimensional dataset embedded within the high-dimensional space, that is to say that there are more descriptors used than objects to describe. The advantage of operating within a subspace is that as well as discarding noise repeated information is merged into a smaller set of descriptors making subsequent computation more robust. The disadvantage is that locating the subspace can be a difficult or computationally expensive task.

2.3 Random Projections

One approach to dimensionality reduction that has seen a recent increase in popularity is the random projection[104], which reduces the dimensionality of the data by projecting the data onto a set of randomly chosen vectors. The Johnson-Lindenstrauss (JL) lemma (Equation 2.1) shows that a set of m points X in a high dimensional space R^N can be embedded into a much lower dimensional space$f : R^N R^k$ such that distance between point are nearly preserved:

$$(1-\epsilon)||u-v||^2 \leq ||f(u)-f(v)||^2 \leq (1+\epsilon)||u-v||^2 \tag{2.1}$$

for all $u, v \exists X$ where

$$k > \frac{8ln(m)}{\epsilon} \tag{2.2}$$

where $0 < \epsilon < 1$. The mapping can be completely random and still achieve this property (see Dasgupta et al [49] for a concise proof). This process removes some data redundancy whilst approximately preserving distances between points, and angles between vectors [17, 49, 134], which is highly desirable for certain types of machine learning and pattern analysis algorithms.

This provides a linear transformation from the original high-dimensional space to a randomly oriented lower-dimensional one. The randomly chosen vectors can be considered analogous to the coefficients (or loadings) familiar from multivariate analysis but as they are not generated in a data dependent manner the whole process is substantially less computationally intensive and thus much faster, this has lead to the specific promotion of its use in environments where computational processing power may be limited [204, 217] but accelerations can be achieved in any environment where large data sets require processing. For example, classification methods that group similar data-points can be used on the randomly projected data as distances are preserved by the projections[68]. Random projections have been used for rapid filtering of secondary ion mass spectra [217] and for data transmission in compressive-projection PCA[69]. Other recent applications of random projections have included work on text mining [17], semantic indexing [133] and classification of gene array data [62, 218]. Classification algorithms typically have computational costs proportional to the number of dimensions [139], so reducing the input dimensionality provides an immediate benefit as well as avoiding the loss of numerical accuracy generally referred to as the *curse of dimensionality*[70].

Compressive sensing is the largest area of application for random projections within the field of hyperspectral imaging[16, 234]. These techniques attempt to construct systems to reduce the number of physical measurements required beneath the Shannon-Nyquist limit. Very recently a theoretical frame work applying compressive sensing to mass spectromety imaging has been developed[15]. Image processing has also benefited from random projections with applications in unmixing[190], clustering[65], face recognition [75] and nearest neighbour finding[17]. In all cases random projections were found to be computationally more efficient with little or no degradation in the quality of results compared to established but computationally intensive methods.

These ideas of random projections were employed within this chapter firstly as a tool for direct reduction of the spectral dimensionality of mass spectrometry images prior to image segmentation. Following this, dimensionality reduction along the spatial dimension is used to produce a domain specific compression model for hyperspectral data. This also results in reduced spectral dimensionality of the data down to approximately that of the subspace occupied by the data. It will be shown how further computations can be performed on the compressed dataset, and how the results can be easily interpreted in the original (physical) measurement

space, e.g. m/z values in mass spectrometry.

2.4 MALDI Imaging of Rodent Brain

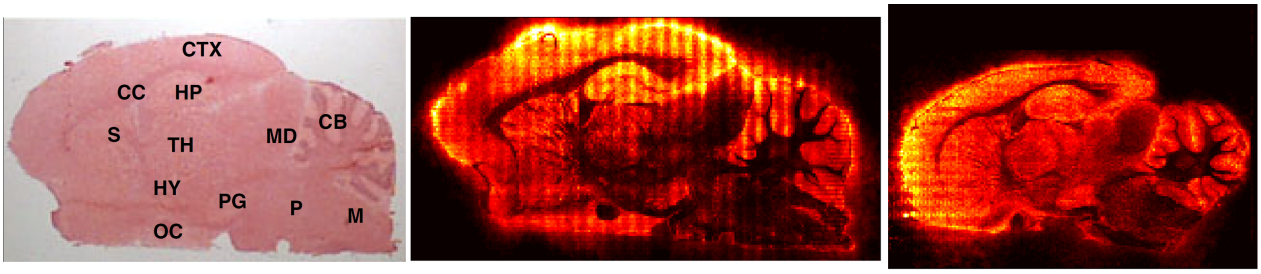

Figure 2.1: Illustration of the MALDI images used for demonstration in this chapter. Images were collected from thin sections of rat brain that was either rapidly snap frozen following excision or formalin fixed then snap frozen. A. H&E stain with key histological structures labelled: cerebellum (cb), cortex (ctx), hippocampus (hi), medulla (med), midbrain (mb), optic chiasm, pituitary gland (pg), pons (p), striatum (st), and thalamus (th). B. Fresh frozen tissue section single exacted ion image showing distribution of lipid PC32:0 [M+K](m/z 771.5). C. Formalin fixed tissue section exacted ion image showing distribution of lipid PC32:0 [M+Na](m/z 756.5).The grid pattern visible is an artefact of matrix deposition. These images were collected by Dr Claire Carter [32] and have been subsequently used for the illustration of MALDI data processing[66, 67, 166]

For the purposes of illustration a single MALDI MSI dataset will be used within this chapter for the illustration of the application of random projections to spectral data. The images, illustrated in Figure 2.1, were collected from a sagittal section of rat brain[32]. A schematic of the images is shown in Figure 2.1 illustrating the anatomical areas of the brain visible in the image area. Note also the 'checkerboarding' effect on the ion images within Figure 2.1 which are believed to be a matrix deposition artefact.

The dataset used here has a dimensionality (number of mass channels) of 129796, and is composed of 20535 individual spectra giving a total data size of ≈20GB. Mass spectra were collected in the small molecule (m/z 50 - 1000) region using an orthogonal quadrupole time-of-flight instrument (QStar Elite QqTOF, AB Sciex, Warrington, UK). The spectra was discretised with 1'count' $\approx$ 100mV and no subsequent data processing was applied. It is fair to assume that there is a substantial degree of spectral degeneracy present within the resulting spectra as mass bins within an individual peak; peaks due to molecular isotopes; and multiple ion-adducts of the same parent molecule will all strongly covary.

This dataset has been used in several papers describing image processing techniques for mass spectrometry imaging [66, 67, 166] and similar regions of brain anatomy were extracted all in cases.

2.5 Orthogonality of Random Projections

Sets of random vectors were generated using the MATLAB (Mathworks, Nant., USA) function `randn`. To verify that the properties required held for large sets of random vectors several sets of differing length were generated and the angle between every pair of vectors was calculated using the dot product:

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}|cos(\theta_{\mathbf{ab}}) \tag{2.3}$$

A histogram of the frequency of each angle was generated for fixed bin widths of 0.2 degrees.

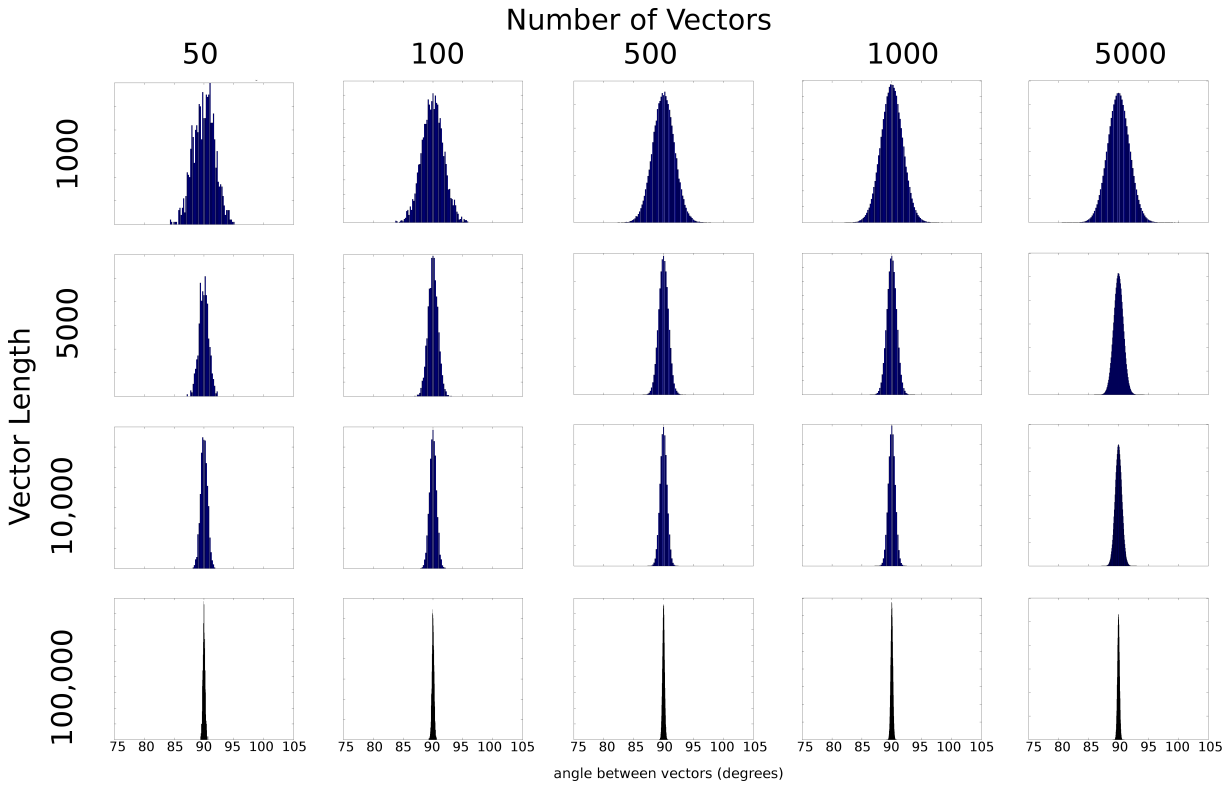

Figure 2.2: The angles between vectors whose elements are randomly drawn from a Gaussian distribution are near-orthogonal. For a set of random vectors, each bar chart tallies the frequency of angle between all vector pairs (tally width 0.2 degrees). Clearly, a typical pair of vectors are orthogonal, as all distributions are centred around 90 degrees, with high probability, as the width of all distributions are narrow.

The property of random projections that makes them so useful for efficiently capturing the information contents of a dataset is the high probability of that any pair of randomly chosen vectors are almost orthogonal i.e. each captures different information about the data. To demonstrate this non-obvious property sets of random vectors (varying from 50-5000) were generated and the angle between each pair calculated. The

number of vectors within the set and the length of each vector was varied and Figure 2.2 shows histograms of the frequency of each angle within sets of different sizes. Each histogram is well localised around 90 degrees with a narrow range (i.e. a typical pair of random vectors is orthogonal). As the length of the random vector increases so does the degree of orthogonality so the projections capture more unique information, and no decrease is seen when a large number of random vectors are compared. From a similar analysis performed by Varmuza et al[217] (on vectors generated from a uniform U[-1 1] distribution) it was suggested that every pair should be evaluated for orthogonality and any sufficiently similar vectors discarded but for the number of measurements in the spectra considered in this work ($\approx 100,000$ m/z values). Figure 2.2 shows this is unnecessary.

Alternatively the set of vectors can be mutually orthogonalised after they are generated [139] at extra computation expense so, conveniently, it makes sense to do this for low vector dimensionality (where the added computational expense is minimal) but is not required for larger dimensional vectors. It is worth noting that these vectors do not form an orthonormal set and so finding the inverse of the random projection matrix (for inverting the projection) cannot be achieved by simple transposition. Dimensionality reduction only provides a computational enhancement if the property required is preserved by the projection (e.g. Euclidean distance, cosine, rank)[139] so care must be taken in the subsequent use of the reduced data. It has been suggested that due to the stochastic nature of random projection ensemble classification can provide more accurate results than a single instance of projections [62, 139].

Having demonstrated the experimental properties of random vectors match the theoretical predictions, experiments with their applications can be performed. The dataset has two directions along which dimensionality reduction could be applied, namely, spectral and spatial. The motivation for applying random projections along each direction will be discussed and results presented.

2.6 Evaluation Metrics

Before the application of random projections are investigated it is useful to establish several evaluation metrics which are useful for understanding the effect of random projections on mass spectrometry imaging data. It will be shown that the usefulness of random projection is that it can be applied to the data without requiring further processing but still enables further analysis. The evaluation is, therefore, always made against the effects that would have been seen with the full data where possible or a reduced version of the original data in the cases where operations are impractical on a full mass spectrometry image.

2.6.1 Absolute Difference

The absolute difference m is the exact difference calculated between the original data $\mathbf{X}$ and a restored version $\mathbf{X}'$. Note that not every projection technique explored will enable a restored version of the data to be produced.

$$m = ||\mathbf{X} - \mathbf{X}'||_2 \tag{2.4}$$

The error metric m measures the absolute ability of a process to represent the data, but it is not a measure of how well specific features of an individual spectrum are preserved.

2.6.2 Signal to Noise Ratio (SNR)

Signal to Noise Ratio (SNR) is commonly used in spectral imaging communities to evaluate the quality of a compression and compares the average numerical deviation between the input data and data that has undergone a compression-decompression cycle. It is computed as

$$\text{SNR} = 10 \log_{10}(\sigma^2/\text{MSE}) \tag{2.5}$$

where

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \bar{\mathbf{X}}_i^T \bar{\mathbf{X}}_i \tag{2.6}$$

and the mean square error of the data points

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{d}_i^T \mathbf{d}_i \tag{2.7}$$

where

$$\mathbf{d}_i = \mathbf{X}_{i:} - \mathbf{Q}\mathbf{X}'_i \tag{2.8}$$

.

2.6.3 Linear Correlation Coefficient

The Pearson's Correlation Coefficient (PCC) is a measure of the linear dependence between an original, $\mathbf{x}$ and restored version of that spectrum, $\bar{\mathbf{x}}$.

$$PCC = \frac{\sum_{i=1}^{m}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}'_i - \bar{\mathbf{x}'})}{\sqrt{\sum_{i=1}^{m}(\mathbf{x}_i - \bar{\mathbf{x}})^2}\sqrt{\sum_{i=1}^{m}(\mathbf{x}'_i - \bar{\mathbf{x}'})^2}} \tag{2.9}$$

The mean value of the vector $\mathbf{x}$ is denoted $\bar{\mathbf{x}}$ and m indexes the m/z bins.

A value close to 1 indicates a strong linear dependence and hence a high quality recovery of major signal features. This can be used both spectrally and spatially to establish whether the information content of the data can be reproduced.

2.7 Spectral Random Projections

The number of spectral channels is usually substantially greater than the number of spectra collected so it seems intuitive to directly project along the spectral domain to reduce the dimensionality. For unsupervised pattern detection any useful dimensionality reduction technique must preserve the chemical differentiation present in the original data. In a tissue image the chemical differentiation would be expected to follow the histology of the tissue section, as different types of tissue have individual molecular signatures.

Algorithm 2.1 details how to spectrally decorrelate using random projections:

Algorithm 2.1: spectral random projection

Data: a spectral image $\mathbf{X}_{m\times n}$ (with m spectral channels and n pixels), number of random vectors, k $(k < m)$

Result: Randomly projected data, $\mathbf{R}_{k\times n}$

1 draw a random matrix $\mathbf{\Omega}_{k\times m}$ from an i.i.d normal distribution $N(0,1)$;

2 Form a new matrix of reduced dimensionality: $\mathbf{R}_{k\times n} = \mathbf{\Omega X}$

As the sampling is random there is no *a priori* way of knowing what chemical information will be captured by a particular projection but by taking many projections all of the spectral information can be captured. Each random projection samples over the whole m/z domain so every projection should capture a unique subset of the chemical information present. From Figure 2.2 it was seen that for vectors of 100,000 elements a very narrow distribution of angles between random projection vectors was seen so, as the spectra in this example have 130,000 elements, there is no need to orthogonalise or sub-select the projection vectors used in this case. To asses whether the random projections preserved chemical differentiation from this mass

spectrometry image a visual assessment of several scores was made.

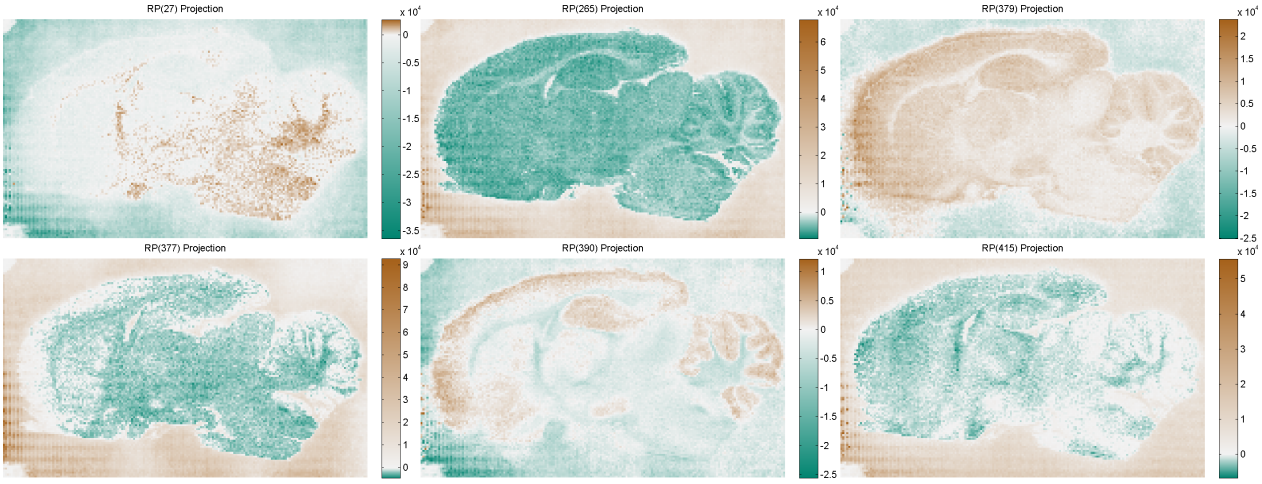

Figure 2.3: Examples of the spatial patterns from random projections strongly suggesting the chemical differentiation is preserved through the process. The distributions visible correspond to the anatomical features visible in Figure 2.1,

The number of spectral channels is usually substantially greater than the number of spectra collected so it seems intuitive to directly project along the spectral domain to reduce the dimensionality.

To asses whether the random projections preserved chemical differentiation from this mass spectrometry image a visual assessment of several scores was made. In a tissue image the chemical differentiation would be expected to follow the histology of the tissue section, as different types of tissue have individual molecular signatures. Figure 2.3 shows six score plots from different random projections of the MSI data, there is clear spatial patterning visible that corresponds to tissue types visible in the schematic and ion image from Figure 2.1. This strongly suggests that molecular differentiation is preserved through the projection.

However, it is not recommended that manual inspection of the loadings is attempted. Individually the projections are unlikely to be directly interpretable due to their random nature but the collection of projections contains the information present within the image. As the RP vectors are chosen from a zero-mean Gaussian they contain some values that are negative, accordingly the scores also have both positive and negative values, this makes physical interpretation of the scores difficult. What can be understood is that each projection corresponds to a random orientation of the projection subspace within the original dimensionality, or a different view of the data. It is worth noting that the random matrix generated changes with every analysis so it is not possible to directly compare projections from separate datasets without pre-generating the random projection vectors.

2.7.1 Preservation of Spectral Magnitudes

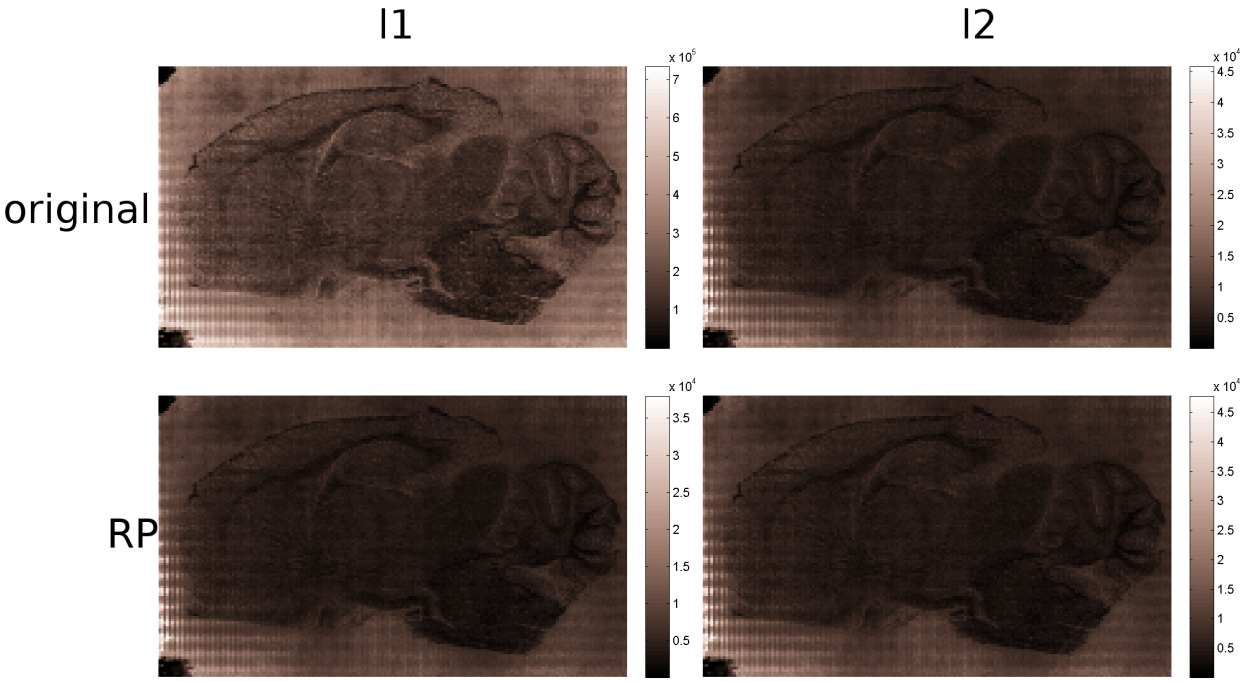

Figure 2.4: Image magnitude recovery from 500 random projections. Top row: the l_1 and l_2 norm calculated for each spectrum in the original data and displayed as an image. Bottom row: the l_1 and l_2 norm calculated for each set of projections and arranged as an image. Comparing the rows, the l_2 norm was successfully recovered but the l_1 norm was not.

The most common vector magnitude measurement encountered in MSI is the l_1 norm applied to spectra which is equal to the 'total ion current' (under the physical assumption that spectra don't contain negative values), further investigations of the use of norms for MSI contained in Chapter 4. Random projections have been shown to preserve the l_2 norm, not the l_1, so direct recovery of the TIC is not possible from the projections, as illustrated in Figure 2.4. The TIC is often used for normalisation of mass spectrometry images, but a comparative study of l_1 and l_2 normalisation for mass spectrometry imaging found little difference in image results using the l_2 norm apart from some bias introduced by its sensitivity towards large magnitude peaks[53].

2.7.2 Unsupervised learning

The preservation of distance metrics by random projections means they can be used for automated data mining, in this case image segmentation. Following projection, the pixels were automatically assigned into clusters using the popular k-means algorithm [2, 108, 210] and each cluster coloured to segment the image. The

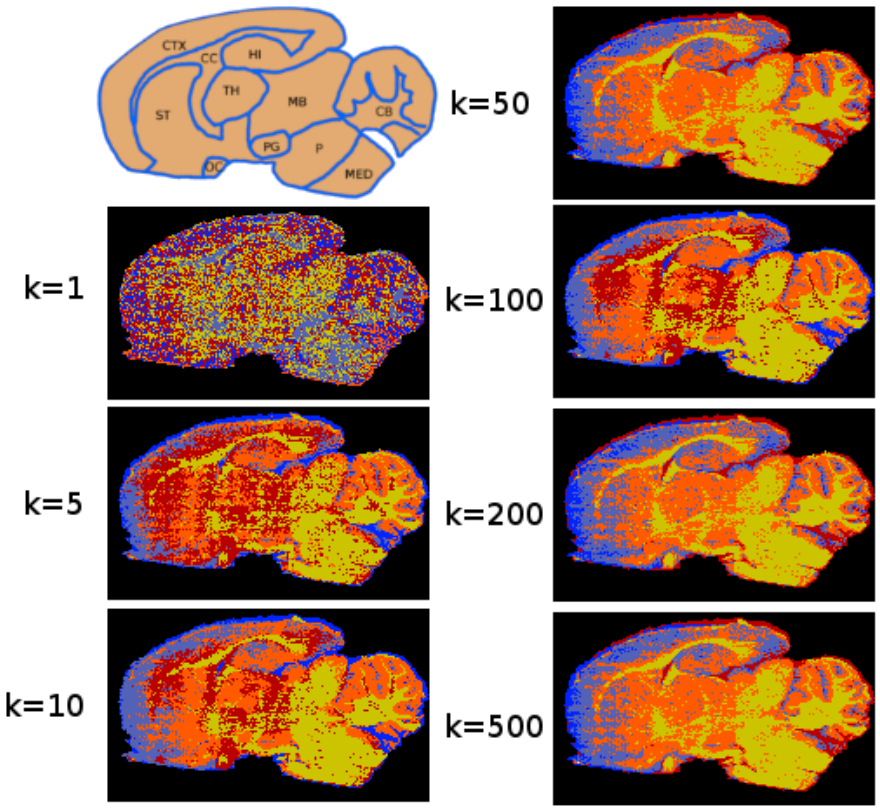

Figure 2.5: Segmenting the fixed rat brain image (into 5 regions using kmeans) as a function of increasing the number of random projections. The clustering extracts the tissue anatomy and becomes stable once 50 or more projections are used

segmentation maps following different numbers of random projections are shown in Figure 2.5. Importantly, this method was applied to a *complete* mass spectrometry image without requiring any of the traditional data processing pipelines of peak detection and feature extraction. The relative spectral characteristics required by this clustering are preserved following projection, but the number of variables to consider for each pixel is massively reduced, providing an immediate computational time saving and avoiding the curse of dimensionality.

A range of numbers of projections were trialled in order to understand the effect that this has on the clustering quality and consistency. To understand the time saving afforded by dimensionality reduction the time required to perform k-means into 5 clusters on each of these projected datasets was recorded, and is show in Figure 2.6. The segmentation algorithm used here, k-means, is linear in the number of dimensions

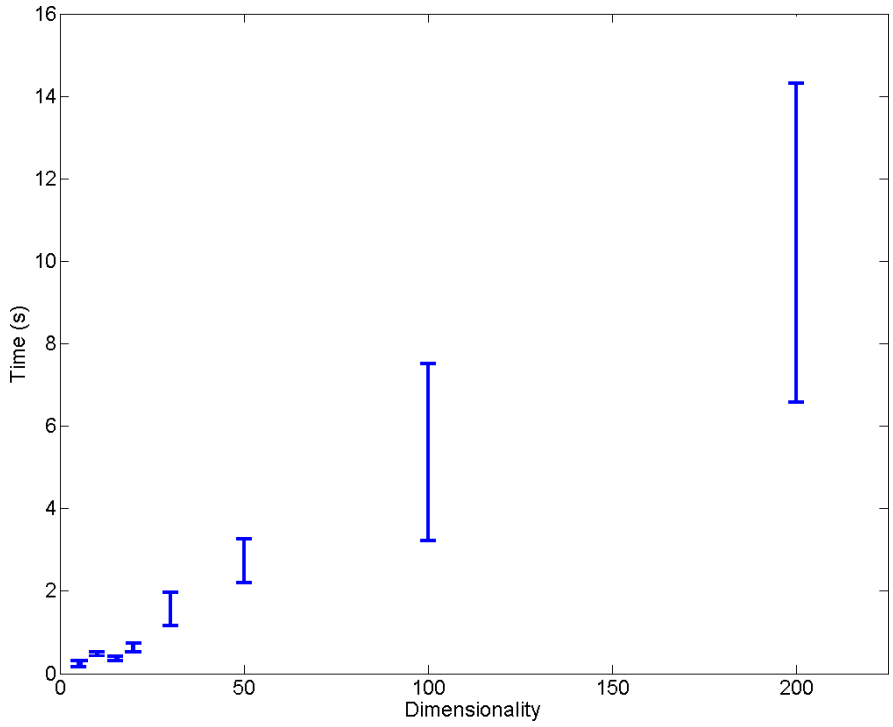

Figure 2.6: Time required for k-means segmentation (5 clusters, error bars show standard deviation of 5 repeats) is approximately a liner a function of the dimensionality

included and this directly translates to the computational speed up. The absolute time required to achieve a stable clustering varies between runs as the algorithm is initialised randomly and so takes different numbers of iterations to stabilise. As expected, the overall time is linear in the number of projections, with some small run-to-run fluctuations. The original data dimensionality was the number of m/z channels, i.e. 130,000, so extrapolating the time required for k-means gives an estimated of 2 hours for the calculations. As the raw

data could not be loaded into memory and would have to be read from disk at each iteration (a process that takes several minutes each time) performing k-means directly on the raw data was not feasible (a single pass algorithm for estimating the k-means centroids exists but is sensitive to the presentation order of the data and requires a second pass for classification[101]).

Segmentation results are shown in Figure 2.5, with a single segmentation map for each choice of number of projections. As discussed in Section 1.6, the most common metric for evaluating the success of segmentation on a tissue image is a comparison with histology. The rodent brain anatomy is well known, as shown in Figure 2.1, and so the segmentation results can be evaluated for the varying numbers of projections. The segmentation maps shown in Figure 2.5 are based purely on the spectral differentiation that is preserved following the projections. A visual comparison to the known anatomy (annotated H&E shown in Figure 2.1 and annotated schematic shown in the first panel of Figure 2.5) shows that the segmentation clearly reproduces the known brain anatomy suggesting that the Mass Spectrometry Imaging (MSI) experiment detected molecular profiles specific to the various anatomic areas.

What is striking is that with only 5 projections (corresponding to a dimensionality reduction of 99.996%) regions are clearly visible. Increasing the number of projections gives a greater stability to the resulting segmentation, and subsequent segmentation maps reveal the same regions. Reproducibility is also increased with greater number of projections, as illustrated in Figure 2.7. In these examples a new set of random vectors is chosen every time before clustering, it is observable that with $k = 5$ the segmentation map produced varies substantially run-to-run whilst with $k = 200$ near identical results are obtained each time. The instability of low numbers of random projections could be offset by using an ensemble approach[65] but this would potentially require the data to be read multiple times which is inefficient when each dataset too large to be stored in main memory.

2.8 Spatial Random Projections

The same concepts of dimensionality reduction introduced earlier can be applied along the spatial dimension of the image to produce a set of spectra that are random combinations of all spectra within the image. This can be achieved using Algorithm 2.2.

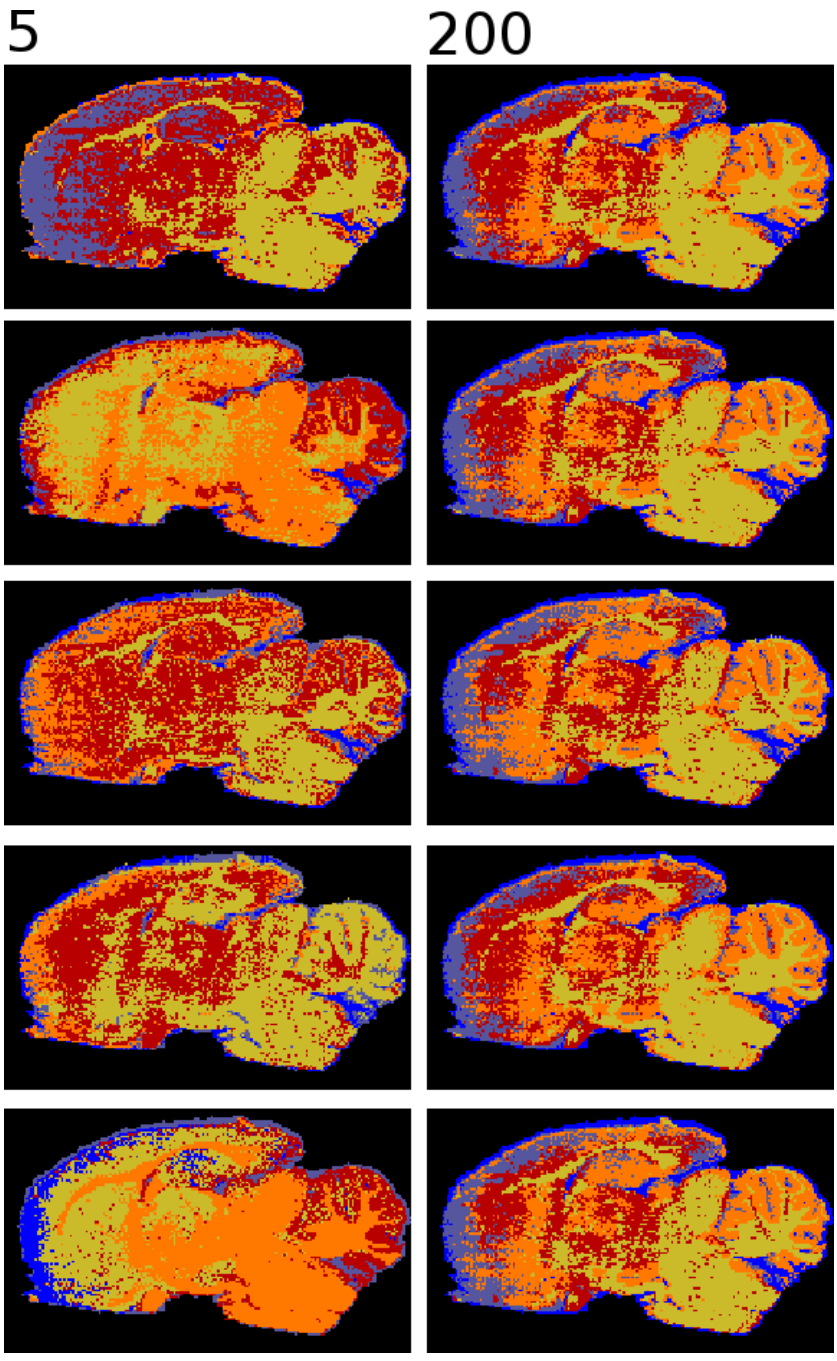

Figure 2.7: Stability of segmentation using k-means increases with greater numbers of projections. The dataset was randomly projected multiple times (with a different set of random vectors generated each time) and segmented using k-means. Left 5 random projections, Right 200 random projections. The variety in segmentation results is much higher for the smaller number of projections.

Algorithm 2.2: Random projection along the spectral dimension

Data: A spectral image$\mathbf{X}_{m\times n}$ (with m spectral channels and n pixels), number of random vectors, k

Result: Randomly projected data, $\mathbf{S}_{m\times k}$

1 draw a random matrix $\mathbf{\Omega}_{n\times k}$ from a normal distribution $N(0,1)$;

2 $\mathbf{S}_{m\times k} = \mathbf{X\Omega}$;

Applying the random projections along the spatial dimension returns a spatially decorrelated set of spectra. Again, achieving the reverse transformation to recover the data in the original space is a non-trivial probabilistic recovery problem.

The main challenge for massive spectral images is that whole dataset cannot be loaded simultaneously and so processing must proceed spectrum-by-spectrum. In practise, this is simple to achieve for projection in both the spatial and spectral directions but as the spatial direction produces a much larger output matrix $\mathbf{S}_{m\times k}$ there is substantially more computational overhead due to the much larger number of matrix multiplication operations required.

Recent work by Halko et al [85] has shown that random projections can be used to construct a low-dimensional orthogonal basis for a dataset, this technique is then applied to spatial projections a spectral basis can be constructed for the data.

2.8.1 Basis Approximation

Hyperspectral imaging data tends to be highly rank-deficient; that is, there are many linear dependencies within the data. This is because channels in the dataset will frequently co-vary with each other, implying that the data exists in a lower-dimensional space than the measurements. If the transformation required to represent the data within this subspace could be found, then the data could be represented and manipulated much more efficiently. In general, this is a computationally expensive task, and the size of many hyperspectral datasets often prohibits a deterministic calculation of the required transformation.

However, it is possible to find this subspace probabilistically. The central idea is that projecting the data onto a set of random vectors preserves (with high probability) the majority of the information within the dataset, provided that the number of random vectors is a little larger than the rank of the full dataset $\mathbf{X}_{m\times n}$ [85]. The random projections can then be orthogonalised (for example, by QR decomposition or Gramm-Schmidt orthogonalisation) to generate an orthonormal basis, $\mathbf{Q}$, of greatly reduced dimension for the data set that preserves most of the information content. A simple procedure for this has been proposed [85] and is described in Algorithm 2.3 and graphically in Figure 2.8:

Algorithm 2.3: Basis approximation to construct an approximate basis for a spectral image[85].

Data: Spectral image, $\mathbf{X}$; integer k

Result: Approximate basis for $\mathbf{X}$, $\mathbf{Q}$

1 Consider a dataset $\mathbf{X}_{m\times n}$ containing n pixels in each of m spectral channels.;

2 Generate random vectors $\left\{\mathbf{v}^{(i)}\right\}_{i=1:k}$ of length n (with $k \gtrsim \text{rank}(\mathbf{X})$) by drawing values from a normal distribution with mean zero and standard deviation one $\mathcal{N}(0,1)$.;

3 Form the random projection matrix $\mathbf{\Omega}_{n\times k} = \left[\mathbf{v}^{(1)}|\ldots|\mathbf{v}^{(k)}\right]$;

4 Project $\mathbf{X}$ onto the random vectors to create a random sampling of the data $\mathbf{S}_{m\times k} = \mathbf{X\Omega}$.;

5 Create orthonormal matrix $\mathbf{Q}_{m\times k}$ from $\mathbf{S}$ by factorising $\mathbf{S} = \mathbf{QR}$, with $\mathbf{R}$ an upper triangular matrix.;

The matrix $\mathbf{Q}$ can be used to project the data onto the lower dimensional subspace and form a reduced data matrix $\mathbf{A}_{k\times n} = \mathbf{Q}^T\mathbf{X}$. This provides a high-quality compressed representation of X provided that a suitable value of the reduced dimensionality k has been chosen. This can be checked by computing the error metric m (using $\mathbf{X}' = \mathbf{QA}$ in Equation 2.4).

The reduced representation is good provided that $m \leq \varepsilon$, for ε a small positive real number, i.e. projecting the data onto the low-rank subspace, and then back-projecting into the original space recreates the original dataset to a high degree of accuracy. In this simple procedure k was selected manually, iterative algorithms exist to find this automatically [85] but are not practical on data stored out of main memory. The consequence of these results is that by storing only the matrices $\mathbf{Q}$ and $\mathbf{A} = \mathbf{Q}^T\mathbf{X}$, the original dataset can recreated via $\mathbf{X} = \mathbf{QA}$. For datasets of low rank, this can be a very considerable saving in memory. The compression ratio is defined as

$$R_c = B_c/B_o \tag{2.10}$$

where where $B_o = m \times n$ is the original data size ($\mathbf{X}$), and $B_c = k(m+n)$ is the combined size of the basis matrix $\mathbf{Q}$ and reduced data matrix ($\mathbf{A}$).

Formalising the compression Model

To formalise the output of Algorithm 2.3, a n-dimensional hyperspectral imaging dataset, $\mathbf{X}_{n\times m}$ is represented using a Basis Approximation for Spectral Compression (BASC) model $\Omega = (\bar{\mathbf{x}}_n, \mathbf{Q}_{n\times p}, m, \mathbf{c}_n)$ as a k-dimensional matix $\mathbf{A}_{p\times n} = \mathbf{Q}^\top\mathbf{X}$, where $\bar{\mathbf{x}}$ is the mean spectrum, $\mathbf{Q}$ is the basis generated using the BASC algorithm [85], m is the number of samples (pixels) and $\mathbf{c}$ stores the channel information (i.e. m/z bins).

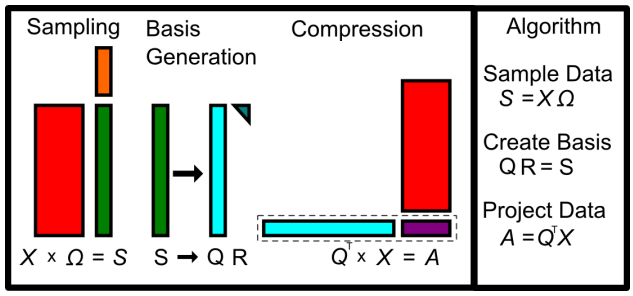

Figure 2.8: Graphical representation of the basis approximation algorithm. The sampling and compression scheme. Form the sampling matrix, $\mathbf{S}$ by multiplying the data, $\mathbf{X}$, by a random matrix, $\mathbf{\Omega}$. An orthonormal basis is formed by QR decomposition of the sampling matrix. Compress $\mathbf{X}$ by onto $\mathbf{Q}$ leaving data stored as a pair of matrices: the basis $\mathbf{Q}$ and scores $\mathbf{A}$.

2.8.2 Basis Approximation for Spectral Compression

The motivation for forming an orthonormal basis for the data is that inverting the projection can be achieved by simply transposing the basis so reduced data $\mathbf{A}$ could be decompressed to form $\mathbf{X}' = \mathbf{QA}$, again, this is demonstrated on the fixed rat brain dataset. The size of the dataset prevented it from being loaded into memory as a whole, so Algorithm 2.4 was modified to allow sequential data access:

Algorithm 2.4: Memory efficient implementation of basis approximation

Data: Spectral image, $\mathbf{X}$; integer k

Result: Approximate basis for $\mathbf{X}$, $\mathbf{Q}$

Consider a dataset $\mathbf{X}_{m\times n}$ containing n pixels in each of m spectral channels.;

Generate random vectors $\left\{\mathbf{v}^{(i)}\right\}_{i=1:k}$ of length n (with $k \gtrsim \text{rank}(\mathbf{X})$) by drawing values from a normal distribution with mean zero and standard deviation one $\mathcal{N}(0,1)$.;

Form the random projection matrix $\mathbf{\Omega}_{n\times k} = \left[\mathbf{v}^{(1)}|\ldots|\mathbf{v}^{(k)}\right]$;

Initialise $\mathbf{S}_{m\times k}$ as a matrix of zeros;

for *j=1* **to** n **do**

  Load jth spectrum from disc $\mathbf{x} = \mathbf{X}_j$;

  Randomly project each spectrum: $\mathbf{S}_{temp} = \mathbf{x\Omega}$;

  Update sampling matrix: $\mathbf{S}_i = \mathbf{S}_{j-1} + \mathbf{S}_{temp}$;

end

Create orthonormal matrix $\mathbf{Q}_{m\times k}$ from $\mathbf{S}$ by factorising $\mathbf{S} = \mathbf{QR}$, with $\mathbf{R}$ an upper triangular matrix.;

Algorithm 2.4 was applied to the rat brain data to generate a randomised basis matrix. The basis matrix produced $\mathbf{Q}$ was used to generate a reduced data matrix through projection $\mathbf{A} = \mathbf{Q}^T\mathbf{X}$ which requires one further pass through data, so only two passes in total. A projection dimension of $k = 100$, producing a

compression ratio of 0.01, is illustrated in Figure 2.9 and was found to adequately represent the data, giving a PCC of greater than 0.99 and an SNR of 45.

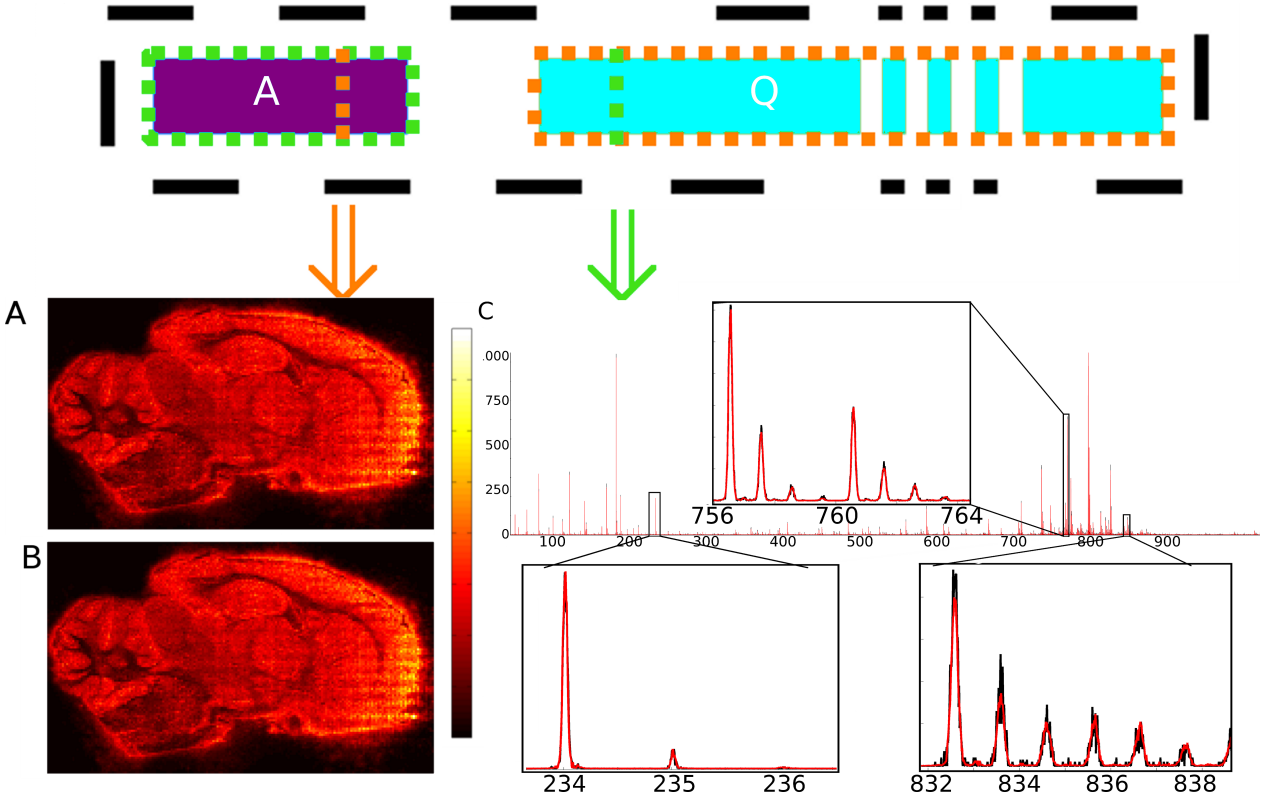

Figure 2.9: Selective decompression of specific spectra and ion images. Decompressed MALDI image. A. Channel map formed from raw data of $m/z = 782.55 \pm 0.05$ showing the distribution of a common lipid, PC(34:1)[32]. B. Channel map showing the distribution following a single compress-decompress cycle of PC(34:1) with $k = 100$. C. Overlay of raw (solid black) and decompressed (red) spectra from a single pixel and there is no substantial deviation following decompression. Enlargements of specific peaks shows that peak shape and intensity are maintained regardless of the initial peak m/z or intensity and the two spectra are still all but indistinguishable.

Qualitatively, this figure shows that the image representation of the selected ion channel is visually indistinguishable from the raw data following a compression-decompression cycle and that decompression of individual spectra show only small differences at the level of the noise. The compression procedure reduced the dataset from the single matrix $\mathbf{X}$ with $m \times n = 129796 \times 20535 = 2665360860$ elements to a pair of matrices $\mathbf{Q}$ and $\mathbf{A}$ with $k(m+n) = 100 \times (129796 + 20535) = 15033100$ elements giving a compression ratio $R_c = 0.0056$. The raw data is $\approx$ 20GB in size, this is reduced to $\approx$ 115MB with this method.

One feature of this compression scheme is that selective decompression can be performed, so decompressing the whole image is avoided when recovering individual spectra or intensity maps, see Figure 2.9 for a visual schematic. Rows in the basis matrix correspond to an individual spectral channel whilst columns in the spatial

matrix correspond to individual pixels. The decompression of a particular single pixel spectrum, i.e. the **j**th *row* from the original data, $\mathbf{X}_j$, can be achieved by selecting the **j**th *column* from the abundance matrix and multiplying it by the basis $\mathbf{X}_j = \mathbf{Q}\mathbf{A}_j$. This is advantageous for minimising the computational memory that needs to be allocated for any operation. The decompression of an intensity map from a particular spectral channel, i.e. a column of the original data $\mathbf{X}_i$, can be achieved by multiplying the appropriate row from the basis and by the whole abundance matrix $\mathbf{X}_i = \mathbf{Q}_i\mathbf{A}$, multiple basis columns are summed row-wise before multiplication for multi-channel images. An image is presented by reforming the resulting list of pixel intensities back to the image dimensions.

To quantify the quality of compression two metrics were calculated: the SNR of the spectra and PCC between the raw and decompressed data. This was done for a range of values of k in order to investigate the trade-off between data size and compression quality. The results of this are shown in Figure 2.10. The SNR and PCC are both seen to be positively correlated with the compression ratio, indicating that taking a large value of k does increase the data quality, but the PCC increases rapidly at first, and flattens out very quickly. The SNR continues to increase, but reaches acceptable values at low compression ratios. In optical hyperspectral imaging SNR values of above 30 for lossy compression are typically considered to be good, and 50 and above excellent[61, 71, 208]. An SNR of 30 is achieved for a compression ratio of < 0.002, corresponding to $k > 35$ on this dataset. For the example in Figure 2.10 with $k = 100$, the SNR is ≈ 43 and the PCC is > 0.99. This suggests that the information lost from the data is at the level of the noise. This curve provides one practical method to estimate a suitable value of k, an alternative starting point is to use the result of the JL lemma (Equation 2.2) and take $k = \frac{8*log(n)}{\epsilon^2}$ (where n is the smallest dimension of the data and is $0 < \epsilon < 1$). However, this approach is known to substantial overestimate the number of projections[17], e.g. for $\epsilon = 0.05$, $k > 30000$.

Comparisons to other mass spectrometry data reduction schemes are difficult as they tend to be based on peak-picking procedures which can be tuned to pick an appropriate number of peaks to compress the data to the size dictated by the computer's memory. Measures of compression quality for the peak-picking methods are not known, but it is commonly accepted that the process of re-binning and peak picking does cause information loss and most efforts have focussed on not discarding 'informative' peaks [66]. Extracting 50-200 peaks has been suggested as appropriate for further analysis such as segmentation [2], which corresponds to a sample-to-feature ratio of around 10 for image sizes typically collected from MALDI time-of-flight experiments. For high-resolution instruments this may require discarding the majority of detectable peaks. The main advantage of the randomised methods is that the dimensionality of the data can be reduced to

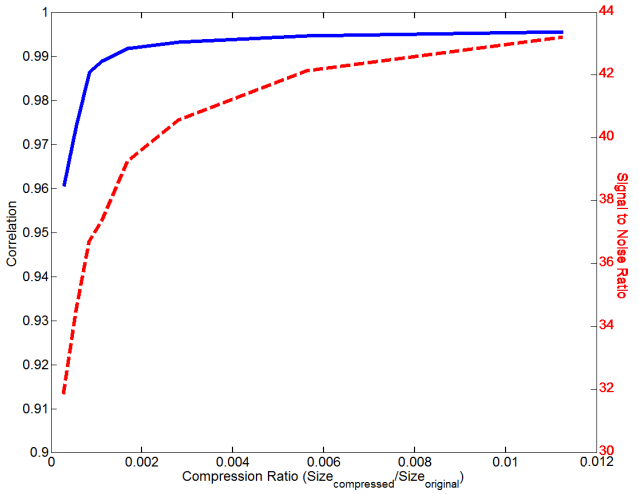

Figure 2.10: Quantitative metrics for evaluating the quality of compression of a MALDI MSI dataset. Pearson product-moment correlation coefficient (solid blue, left axis) and signal-to-noise ratio (dashed red, right axis). As values of PCC tend to one a high quality signal recovery is achieved.

a similar level as that obtained by peak-picking methods, but no part of the data is removed completely. Further, the dimensionality reduction obtained via these means is reversible. The discrete wavelet transform, a more comparable method, has been employed in an attempt to preserve the spectral integrity of the data [215]. This was shown to reduce the dimensionality from 6490 to 819 but details of total data size and metrics of compression quality were not presented. On this limited basis, randomised basis approximation appears to be able to achieve superior compression ratio to wavelet-based methods whilst maintaining the quality of the data.

Image Magnitude Recovery

Like the direct random projection the basis approximation projection preserves the l_2 norm but not the l_1, see Figure 2.11, however, the ability to back-project to the original data means that the l_1 norm is not entirely lost through the process. It can be recovered without requiring the full dataset be back-projected by taking the sum of each basis vector and back-projecting.

$$l_{1n} \approx \left(\sum_{i=1}^{m} \mathbf{Q}_n \right) \cdot \mathbf{A} \tag{2.11}$$

Taking a similar approach with random projection, see Figure 2.12 does not yield such results. This is particularly useful as the TIC is commonly used for image normalisation.

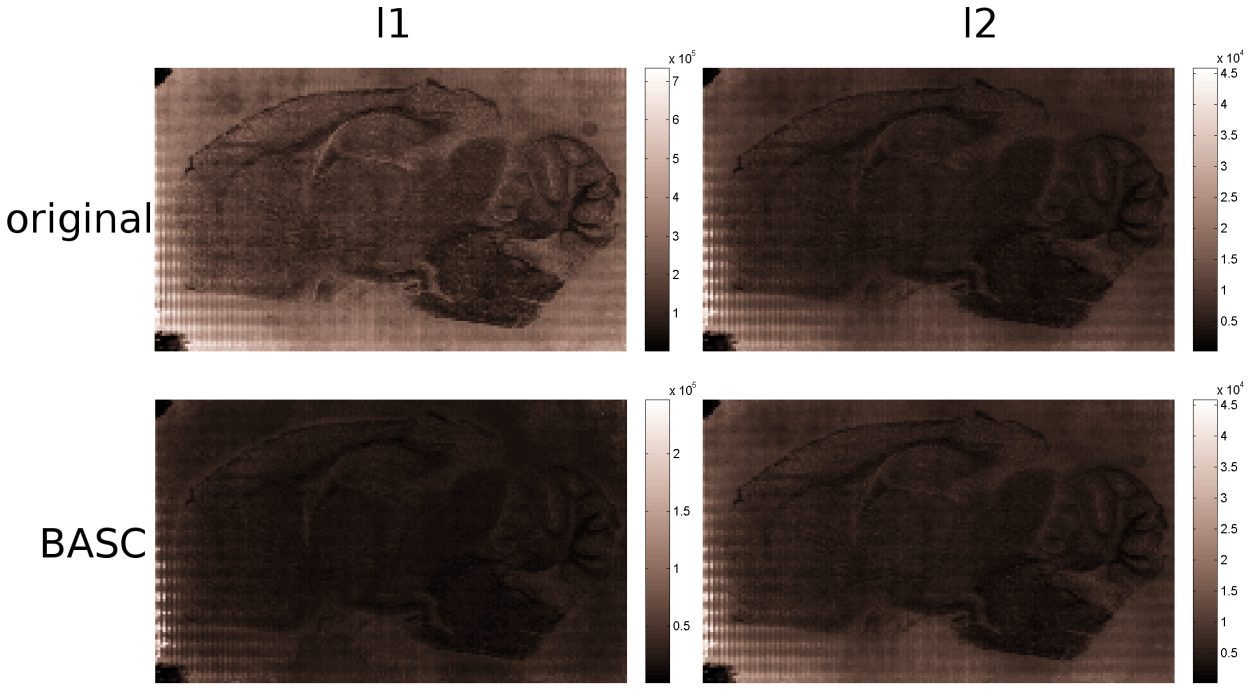

Figure 2.11: Recovery of data norms from data compressed with BASC. It was not possible to recover the l_1 norm (as calculated directly on the data) from the BASC projections but the l_2 norm can be calculated directly from the BASC projections

2.8.3 Pixel Sub-Sampling

The procedure for computing a randomised basis can be further refined by the following observation. The computation forms random linear combinations of the data points, and there is very likely to be some degree of homogeneity in the data points, since, for example, neighbouring pixels are likely to have very similar spectra. It is therefore not necessary to include all data points in the basis computation, and the data can be *spatially randomly subsampled* to increase the efficiency of the computation as explained in Algorithm 2.5. The basis generated from a randomly chosen subset of the data will, with high probability, form a basis for the full dataset as long as the sampling density is sufficient. This can easily be checked by ensuring that the error metric $M < \varepsilon$. The use of spatial sub-sampling reduces the amount of computation that must be performed in order to compute a random basis for the dataset and is particularly advantageous on datasets with many pixels and a high degree of spatial homogeneity.

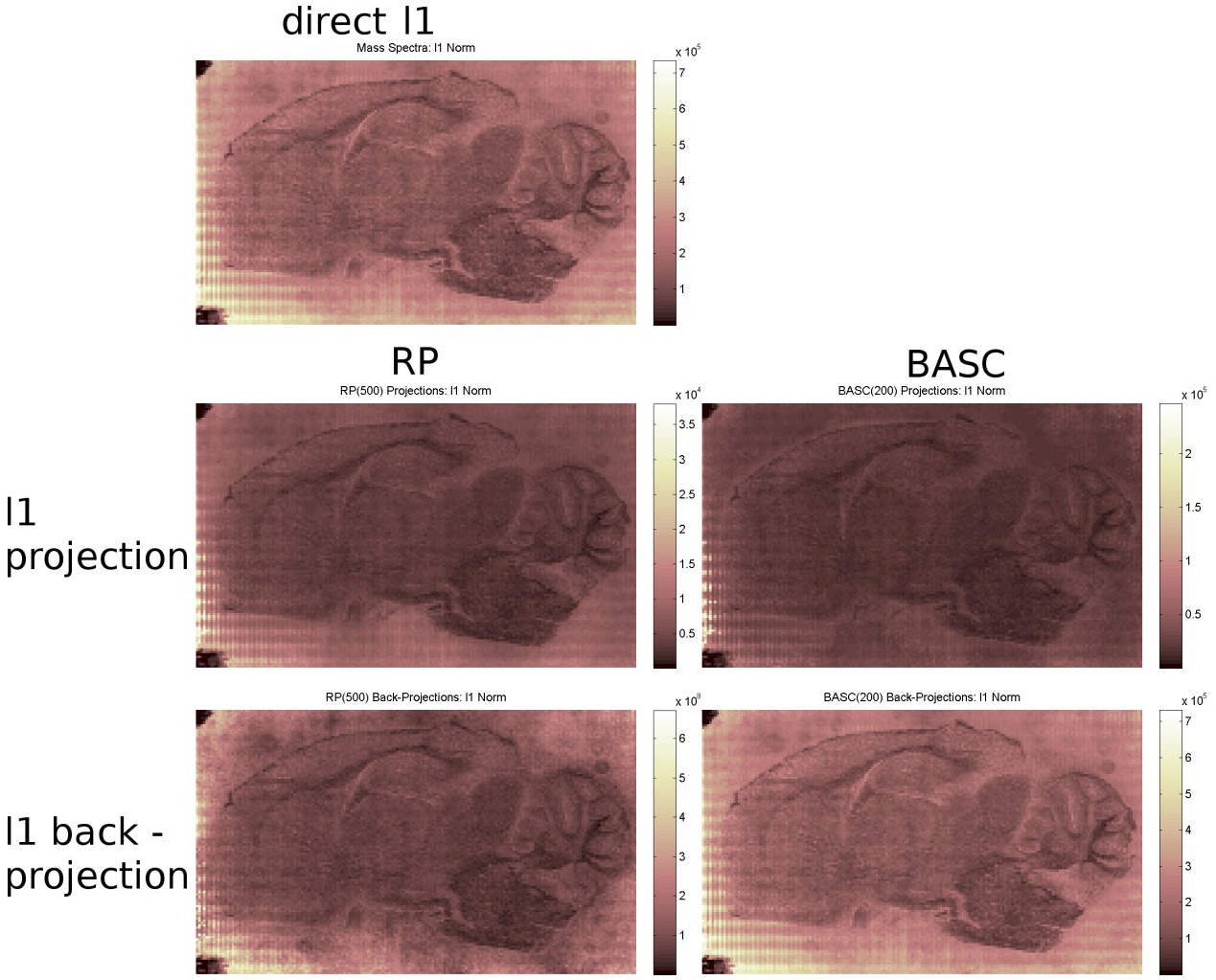

Figure 2.12: BASC compressed data can be back-projected to recover the l_1 norm for a mass spectrometry image.

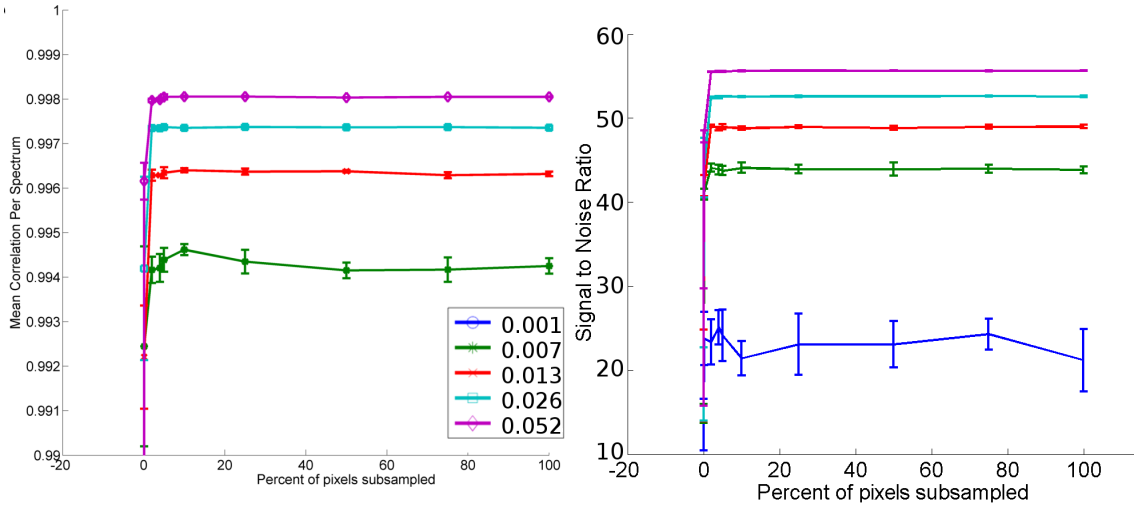

Figure 2.13: Analysis of compression quality obtained using spatial sub-sampling from the reduced MALDI mass spectrometry dataset with a range of compression ratios. Left: Pearson correlation coefficient. Right: Signal to noise ratio. A random chosen subset of pixels was used in every trial. Error bars show variation (one standard deviation) from 10 trials. Both the PCC and SNR stabilise after the sub-sampling exceeds 10% regardless of the compression ratio, again a higher compression ratio provides improvement in compression quality.

Algorithm 2.5: Generate an approximate basis for a spectral image with spatial subsampling

Data: Spectral image, $\mathbf{X}$; integer k; integer r

Result: Approximate basis for $\mathbf{X}$, $\mathbf{Q}$

1 Consider a dataset $\mathbf{X}_{m\times n}$ containing n pixels in each of m spectral channels.;

2 Randomly select r columns from $\mathbf{X}$ to form a spatially sub-sampled data matrix $\mathbf{Y}_{m\times r}$;

3 Generate random vectors $\left\{\mathbf{v}^{(i)}\right\}_{i=1:k}$ of length r (with $k \gtrsim \text{rank}(\mathbf{X})$) by drawing values from a normal distribution with mean zero and standard deviation one $\mathcal{N}(0,1)$.;

4 Form the random projection matrix $\mathbf{\Omega}_{r\times k} = \left[\mathbf{v}^{(1)}|\ldots|\mathbf{v}^{(k)}\right]$;

5 Project $\mathbf{Y}$ onto the random vectors to create a sampling matrix $\mathbf{S}_{m\times k} = \mathbf{Y}\mathbf{\Omega}$.;

6 Create orthonormal matrix $\mathbf{Q}_{m\times k}$ from $\mathbf{S}$ by factorising $\mathbf{S} = \mathbf{QR}$, with $\mathbf{R}$ an upper triangular matrix. ;

7 The compression is then achieved by projecting $\mathbf{X}$ onto $\mathbf{Q}$ to form $\mathbf{A} = \mathbf{Q}^T\mathbf{X}$.

Performance Improvements with Pixel Sub-Sampling

A performance improving algorithm for basis approximation is described in Algorithm 2.5 which takes only a small subset of the data to use for generating the reduced basis. This method expects there to be some spatial homogeneity within the image area in order to reduce the number of spectra sampled to span the subspace occupied by the data. A range of spatial sub-sampling rates and compression ratios were investigated using

Algorithm 2.5 and the PCC and SNR error measures to asses the compression quality produced by spatially subsampling when generating the randomised basis. Spatial points were randomly chosen for producing the basis, Figure 2.13 shows the error metrics. The compression quality stabilises at a sampling rate of less than 20% in all cases. The compression ratio has little effect on the sampling rate needed, but improves the quality of compression independently. These results are sample-specific, being dependent on the degree of spatial homogeneity in the sample.

It is therefore not necessary to access the entire dataset in order to generate the randomised basis. This is likely to significantly reduce the time required to process very large datasets where disc access time is a significant cost. It will, of course, still be necessary to read the whole dataset in order to perform the projection, but the use of random sampling means that the whole data need only be read in once.

2.8.4 Forming a Mutual Basis for Multiple Datasets

A single MSI dataset may not be sufficient to answer a biological question, for example in the comparison of tissue collected from cohorts of normal and diseased subjects, and so multiple images may be acquired. Statistical analysis over several images requires that the data be combined into a single measurement matrix. It is possible to analyse individual ion signals by loading each dataset independently but even in this case it would be preferable to be able to view all images at once. Combining datasets for analysis is equivalent to concatenating the individual measurement matrices, so the objective is to merge two separate MSI datasets,$\mathbf{X}$ & $\mathbf{Y}$, to form a single data object: $\mathbf{Z}_{n\times t} = [\mathbf{X}, \mathbf{Y}]$. The combined dataset will then be compressed using a BASC model (as defined in Section 2.8.1) for further analysis. It is assumed that $\mathbf{X}$ & $\mathbf{Y}$ have already been compressed and so have BASC models Ω and Ψ respectively.

Naively this could be achieved calculating a basis representation by returning to the original data (either the raw data or by completely decompressing the compressed representations) and calculating a mutual basis for the concatenated data. However, as the spectral and spatial variation of each dataset is contained within the BASC models Ω & Ψ this concatenation can be achieved without re-inflating the data. This could be achieved by simply concatenating the two bases but for most datasets being compared there would be expected to be some overlap in the spectral content. Shared spectral information should be concentrated in the final basis to minimise the number of basis vectors that are required to describe the combined data. The procedure for merging two BASC models to find a combined model Φ using only the models Ω and Ψ. is provided in Algorithm 2.6.

Algorithm 2.6: Find a BASC model, $\Phi = (\bar{\mathbf{z}}_{n\times 1}, \mathbf{T}_{n\times v}, t, \mathbf{c}_{n\times 1})$, for two concatenated input datasets $[\mathbf{X}, \mathbf{Y}]$

Data: two hyperspectral images $\mathbf{X}_{m\times n}$, $\mathbf{Y}_{q\times n}$, merge threshold ϵ

Result: BASC model, $\Phi = (\bar{\mathbf{z}}_{n\times 1}, \mathbf{T}_{n\times v}, t, \mathbf{c}_{n\times 1})$

1 Form BASC models for the two hyperspectral images $\mathbf{X}$ and $\mathbf{Y}$ using Algorithm 2.4, respectively, $\Omega = (\bar{\mathbf{x}}_{n\times 1}, \mathbf{Q}_{n\times p}, m, \mathbf{c}_{n\times 1})$ and $\Psi = (\bar{\mathbf{y}}_{n\times 1}, \mathbf{R}_{n\times s}, q, \mathbf{c}_{n\times 1})$;

2 By definition $t = m + q$ and so the new mean is $\bar{\mathbf{z}} = \frac{1}{t}(m\bar{\mathbf{x}} + q\bar{\mathbf{y}})$;

3 Calculate whether any portions of R already overlap with Q by calculating the residual: $\mathbf{H}_{n\times u} = \mathbf{R} - \mathbf{Q}\mathbf{Q}^\top\mathbf{R}$;

4 Some of the column vectors within $\mathbf{H}$ are likely to represent overlap between Ω and Φ and so vectors with l_2 norms smaller than ϵ can be remove to form a reduced matrix $\mathbf{H}'_{n\times u'}$.;

5 The residual is also calculated of a vector joining the means with respect to Ω to accommodate for space between datasets to be included in the final basis space: $\mathbf{h} = (\bar{\mathbf{x}} - \bar{\mathbf{y}}) - \mathbf{Q}\mathbf{Q}^\top(\bar{\mathbf{x}} - \bar{\mathbf{y}})$;

6 An orthonormal basis spanning the additional information, which needs to be included in the final basis,$\mathbf{V}_{n\times u'+1}$, can be obtained, e.g. using qr decomposition of $[\mathbf{H}, \mathbf{h}]$.;

7 The combined basis $T_{n\times v}$ is formed: $\mathbf{T}_{n\times v} = [\mathbf{Q}_{n\times p}\mathbf{V}_{n\times u'+1}]$ where $v = p + u' + 1$.

Allowing for an expansion in the number of basis vectors to encompass additional information content so $\mathbf{T}$ has a size bounded by $max(p, q) \leq v \leq p + s + 1$ where the $+1$ arises from the addition of the vector joining the data centres.

Notice that both datasets, $\mathbf{X}$ & $\mathbf{Y}$, and thus models, Σ & Φ, have consistent channel information, $\mathbf{c}$ (i.e. the datasets are collected over the same spectral region). In general this may not be the case for two datasets but is required for basis merging.

Modelling the Data on the Mutual Basis

Another pleasant result of this compression scheme is that the data can be transferred to the new model without having to be restored to its decompressed size.

Algorithm 2.7: Project two image datasets onto a single basis

Data: two hyperspectral images $\mathbf{X}$, $\mathbf{Y}$ with their BASC models Ω and Ψ, mutual BASC model Φ

Result: $\mathbf{C}$, dataset compressed with BASC model Φ

1 Project $\mathbf{X}$ & $\mathbf{Y}$ using the bases within their respective models:

$$\mathbf{A}_{p\times n} = \mathbf{Q}^\top \mathbf{X} \tag{2.12}$$

$$\mathbf{B}_{s\times n} = \mathbf{R}^\top \mathbf{Y} \tag{2.13}$$

In general, our data is represented in Ψ as

$$\mathbf{C}_{v\times t} = \mathbf{T}^\top \mathbf{Z} \tag{2.14}$$

2 as $\mathbf{Z} = [\mathbf{X}, \mathbf{Y}]$ and from BASC algorithm there are approximate matrix equations for $\mathbf{X}$ and $\mathbf{Y}$ (Equations 2.12 & 2.13) using these Equation 2.14 can be rewritten as:

$$\mathbf{C} = [(\mathbf{T}^\top \mathbf{Q})\mathbf{A}, (\mathbf{T}^\top \mathbf{R})\mathbf{B}] \tag{2.15}$$

So the compressed data can be projected directly onto the new basis without having to re-inflate to the original size, with the small error that existed in each individual projection and some round-off error from the basis combination. Multiple compressed datasets could be combined by applying this algorithm iteratively. Many data reduction schemes in do not allow the addition of subsequent data without starting the data processing routine from scratch with all the data in its pre-compressed form so having this ability extends the usefulness of this approach.

Demonstration with a Single Dataset

To demonstrate the merging process a single dataset was divided into two halves and recombined, as illustrated in Figure 2.14. The fixed brain dataset was a)compressed whole using BASC b)split in half (spatially), each half was compressed independently using BASC, the two halves were then combined. In every case 200 basis vectors were constructed. Each half was compressed independently and subsequently merged, without having to access the raw data again. The projection matrix produced for each half was different and each image would be expected to have a different molecular profile. Nevertheless, upon dataset merging no evidence of a divide was noted between image halves and identical principal component analysis results were obtained

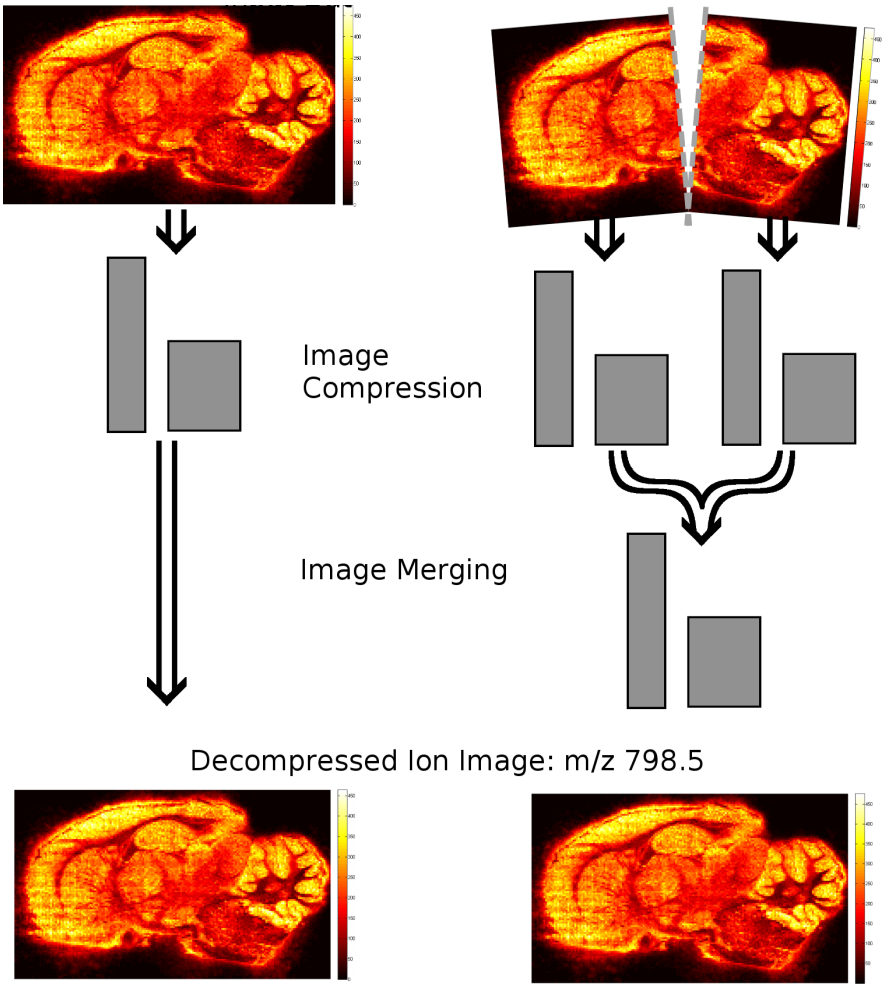

Figure 2.14: Illustration of the capabilities of dataset merging. A single dataset is processed whole (left), and as two halves (right) (k = 200). The two half images are compressed independently (producing two basis sets) and then recombined (whilst compressed) to form a single compressed image (k=401). Data analysis can be performed on the compressed images, such as ion image viewing (m/z 798.55±0.05), showing that data information integrity is maintained during the merging process.

as from the intact dataset. An SNR of 42.5 was measured for the merged data (an SNR of 42.5 was also produced for BASC on the full dataset) showing that a high performance was achieved by the basis merging, identical to that of direct compression.

One obvious application of merging basis is to facilitate the comparison of two datasets that have been independently compressed. The main obstacle in this case is ensuring that the same m/z domain exists for both sets, which can be different both in range and in bin spacing. A difference in range can only be dealt with by cropping to the overlap region, so just the mutual m/z range is considered. If the m/z binning is different then the two datasets cannot be directly combined, as the basis is indexed against a known m/z scale, and so must be interpolated onto a common axis first. Overall intensity differences may be observed between the images and so pixel-wise normalisation may be required to enable comparison.

Comparing Fresh and Fixed Brain Images The datasets introduced in Section 2.4 were produced from tissue that had undergone different fixation steps during tissue collection and it was shown by Carter et al[32] that this resulted in a change in the predominant metal adduct seen from potassium to sodium. Reproducing these results provides a useful example of the the kinds of analysis that are now facilitated by being able to merge bases. Formally, to make the comparison of species detected between two datasets ion-images corresponding to the molecules considered would have been made independently for each dataset and then aligned and compared. After merging these two datasets using Algorithm 2.6 a mutual basis was calculated using a threshold of 10, the datasets were then combined using Algorithm 2.7. The same ion image can then be produced simultaneously from both datasets for immediate comparison, producing results that are qualitatively identical to generating the ion images independently, see Figure 2.15.

2.9 Conclusions

Random projections have been shown to have significant utility in the processing of mass spectrometry images. When used directly to address the spectral dimensionality issue they were found to provide effective dimensionality reduction. By using a data-independent pseudo-basis in a memory efficient manner all of the data reduction could be performed in a single stage, significantly reducing the complexity and time of MSI data processing pipelines. The specific choice of random distribution to use has been addressed in the literature and several computationally faster sets have been proposed [1] that also match the properties of a normal distribution. In this thesis all random vectors will be chosen from a normal distribution, leaving the potential for further speed gains during a full software deployment.

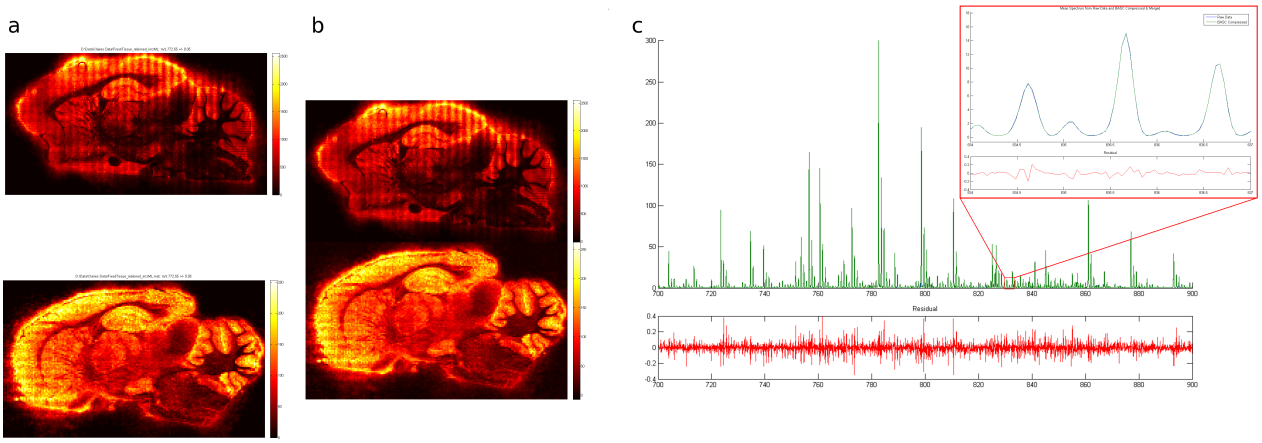

Figure 2.15: Merging two BASC compressed MALDI MSI datasets. a) Example raw ion images from m/z 772.5±0.05 produced separately from the datasets. b) The same ion images reproduced after BASC compression and dataset merging (image-subsections are colour scaled separately for comparison with *a*) c) Overlay of mean spectrum from raw (blue) and compressed data (green). The plot of the difference between the two (red) shows quantitatively mean spectral features are preserved down to the level of the data noise, even for very small peaks.

Applying random projections to decorrelate the spatial direction of MSI data lead to the construction of an approximate basis for the data with a single pass of the data. Projection onto the basis required a second pass through the data but due to the high levels of data preservation the basis and projection could then be used as a data compression scheme. Individual spectra and ion images could be recovered following projection and datasets could be combined. The use of this technique for data compression provides useful compression ratios combined with the advantage of being able to continue mining the data whilst compressed. It is not a lossless process but stationary noise is not preserved well by this approach so the compression actually de-noises the data slightly. Combining datasets will be very powerful for large imaging datasets such as those collected by 3D images, cohort studies or data streaming applications.

The choice of whether to use basis approximation over random projection should be made on the number of processes that are likely to be performed on the data. If comparisons with other datasets are likely, or several segmentations may be required then the extra time required to build the basis will be compensated for as no further passes through the full data are required. The extensibility of the techniques developed to other modalities will be explored in Chapter 6 but it should be noted that nothing is assumed of the data other than its spectral nature.

2.10 Acknowledgements

I am most grateful to Dr. Claire Carter for sharing the rat brain data used within this, and subsequent, chapters.

Chapter 3

Spectral Data Mining after Compression with Basis Approximation

---

Basis approximation was shown to provide a representation of the data that reduced the total data volume whilst preserving the spectral and spatial information contained. As a linear transformation it also preserves the l_2 norm as well as linear relationships within the data. This preserves much of the structure of the data during the compression process and this enables a substantial amount of data analysis to be performed on the data *without decompression* providing a completely new compressed analysis approach for mass spectrometry imaging.

In this Chapter the application of segmentation and factorisation by both Principal Component Analysis (PCA) and Non-negative matrix factorisation (NNMF) to images compressed with BASC was investigated and, where it was possible to compare, equivalent results are obtained to analysing the raw data but with substantially reduced computation time.

3.1 Introduction

The two most common approaches to unsupervised data mining of Mass Spectrometry Imaging (MSI) are factorisation and segmentation, the principles and applications of which were discussed in Chapter 1. Both aim to automatically visualise or label the samples by identifying spectral trends that differentiate between pixels. A limitation when applying both of these methods, and indeed probably any machine learning technique, to MSI is the large dimensionality of the data. This imposes limits, also discussed in Chapter 1, on the application of unsupervised analysis due to practical computational considerations and the 'curse of dimensionality'.

Basis approximation was established in Chapter 2 as a statistically sound technique for finding an orthonormal basis for spectral images allowing their compression and decompression with no losses above the noise baseline. This basis-projection model was also shown to be able a useful way for combining separate datasets for comparison and in all cases produced a substantial dimensionality reduction. These results agreed with work by Halko et al[85] that demonstrated that the basis found through this randomised approach retains nearly all of the information present in the original matrix.

The basis produced by the randomised method consists of a set of linearly independent basis vectors, so the projection onto the basis preserves linear relationships within the data. Factorisations, such as PCA or NNMF, also seek linear trends within the data to form a basis from. As these linear trends are preserved during the projection onto the randomised basis the factorisation algorithms can be applied to the projected data. This has been demonstrated for singular value decomposition (Singular Value Decomposition (SVD))[85], and so this thesis will develop similar methods for PCA and NNMF. The preservation of distance during projection is also important for segmentation, as was utilised in Chapter 2, and basis approximation accurately preserves relative distances and angles (and therefore inner products)[85]. This allows the distance calculation to be performed on a reduced, and concentrated, dimensionality and so providing faster more accurate results.

The main advantage of computing the factorisation on the projected data is that no data reduction step which explicitly removes or down-samples mass to charge ratio (m/z) channels is required so the results should be equivalent to computing on the raw data. Due to the reduced dimensionality the algorithms also take much less time to compute. Furthermore, the compression scheme allows for simple and high quality decompression back into the original space from the factors or cluster centroids located within the compressed data, preserving the molecular information as well as spatial patterns.

This chapter examines the suitability of several factorisation and clustering algorithms for use with data

compressed using basis approximation and considers some of the challenges of visualising the outputs. The performance of these algorithms on compressed data will be evaluated against their operation directly on the mass spectrometry data (that has had its spectral dimension substantially reduced) to demonstrate that equal results are obtained. The effect of altering the compression amount (by choosing fewer basis vectors) will be explored and the patterns extracted using these techniques are compared to the biology of a well characterised sample.

3.2 Compressed Factorisation

As the most popular factorisation method applied to mass spectrometry data, it will first be considered how PCA can be performed on a compressed dataset.

3.2.1 Principal Component Analysis (PCA)

In general, PCA is performed by computing the eigenvalues and eigenvectors of the covariance matrix but given the reduced dataset $\mathbf{A}_{k\times n}$ the eigenvalue decomposition of the compressed covariance matrix can be computed and the eigenvectors subsequently decompressed. In the literature the eigenvectors are often referred to as PCA coefficients or PCA loadings, the term coefficients are used here but all three are interchangeable in this context. First, the basic PCA calculation is reviewed.

The use of an orthogonalised random projection has two important consequences. Firstly, it reduces the dimensionality of the problem. Using this, it is possible to perform PCA on the reduced data matrix $\mathbf{A}$ instead of on the full data matrix $\mathbf{X}$. On a full dataset, this calculation requires the formation and diagonalisation of the $m \times m$ covariance matrix, with $m \approx 10^5$ for Matrix Assisted Laser Desorption Ionisation (MALDI) datasets. This is intractable by normal means, although special data reduction techniques can be employed in combination with memory-efficient algorithms which construct the PCs without requiring the data to be in memory [172]. Here, it is only required that the $k \times k$ covariance matrix from the reduced subspace is diagonalised, from which no data channels have been fully removed. Secondly, the orthogonalisation procedure will allow the principal component vectors to be projected back into the full high-dimensional measurement space so that they can be analysed in terms of physically meaningful quantities. This is the key advantage of this approach over non-orthogonalised random projection methods.

Algorithm 3.1: Compute the principal component eigenvectors and eigenvalues from a data matrix

Data: A data matrix, $\mathbf{X}$

Result: eigenvectors,$\mathbf{W}$; eigenvalues, $\mathbf{\Phi}$

1 Given a data matrix $\mathbf{X}_{m\times n}$, compute the mean of each row (data channel) and subtract to form matrix $\bar{\mathbf{X}}$ such that each row of $\bar{\mathbf{X}}$ has zero-mean.;

2 Compute the covariance matrix $\mathbf{C} = \bar{\mathbf{X}}\bar{\mathbf{X}}^T$;

3 Calculate the eigenvalue/vectors of $\mathbf{C}$. Noting that $\mathbf{C}$ is a symmetric real matrix, this is equivalent to performing SVD on $\mathbf{C}$: $\mathbf{C} = \mathbf{W}\mathbf{\Phi}\mathbf{W}^T$, The columns of $\mathbf{W}$ are the principal component vectors, and the diagonal of $\mathbf{\Phi}$ contains the variance along each of the principal components.;

The use of an orthogonalised random projection has two important consequences. Firstly, it has reduced the dimensionality of the problem. PCA can be performed on the reduced data matrix $\mathbf{A}$ instead of on the full data matrix $\mathbf{X}$. On a full dataset, this calculation requires the formation and diagonalisation of the $m \times m$ covariance matrix, with $m \approx 10^5$ for MALDI datasets. This is intractable by normal means, although special data reduction techniques can be employed in combination with memory-efficient algorithms which construct the PCs without requiring the data to be in memory[172]. It is enough to simply diagonalise the $k \times k$ covariance matrix from the reduced subspace, from which no data channels have been fully removed. Secondly, the orthogonalisation procedure allows the principal component vectors to be projected back into the full high-dimensional measurement space so that they can be analysed in terms of physically meaningful quantities. This is the key advantage of this approach over non-orthogonalised random projection methods.

It should be noted that an alternative algorithm exists for obtaining the eigenvectors using SVD factorisation of the data matrix directly, rather than explicitly forming the covariance matrix[105] which also suffers due to the data size. Identical arguments can be made for applying this method to the compressed data, and the same transformations can be used to recover the principal components in the original space, as detailed in the following Algorithm.

Algorithm 3.2: Compute the principal component eigenvectors and eigenvalues from a compressed data matrix

Data: Compressed data matrix, $\mathbf{A}$; and its approximate basis, $\mathbf{Q}$

Result: eigenvectors,$\mathbf{W}$; eigenvalues, $\mathbf{\Phi}$

1 Form the zero-mean reduced data matrix $\bar{\mathbf{A}}$ using Algorithms 2.4 and 3.1.1. ;

2 Form the covariance matrix in the reduced subspace, $\tilde{\mathbf{C}} = \bar{\mathbf{A}}\bar{\mathbf{A}}^T$.;

3 Perform SVD to diagonalise: $\tilde{\mathbf{C}} = \tilde{\mathbf{W}}\tilde{\mathbf{\Phi}}\tilde{\mathbf{W}}^T$, giving the principal components and variances in the reduced subspace.;

4 Since $\bar{\mathbf{A}} = \mathbf{Q}^T\bar{\mathbf{X}}$, $\tilde{\mathbf{C}} = \bar{\mathbf{A}}\bar{\mathbf{A}}^T = \mathbf{Q}^T\bar{\mathbf{X}}\bar{\mathbf{X}}^T\mathbf{Q}$.;

5 It then follows that $\mathbf{Q}\tilde{\mathbf{C}}\mathbf{Q}^T = \mathbf{Q}\mathbf{Q}^T\bar{\mathbf{X}}\bar{\mathbf{X}}^T\mathbf{Q}\mathbf{Q}^T \approx \bar{\mathbf{X}}\bar{\mathbf{X}}^T = \mathbf{C}$ As $\mathbf{C} = \mathbf{W}\mathbf{\Phi}\mathbf{W}^T$ and $\tilde{\mathbf{C}} = \tilde{\mathbf{W}}\tilde{\mathbf{\Phi}}\tilde{\mathbf{W}}^T$ it follows that $\mathbf{W}\mathbf{\Phi}\mathbf{W}^T = \mathbf{Q}\tilde{\mathbf{W}}\tilde{\mathbf{\Phi}}\tilde{\mathbf{W}}^T\mathbf{Q}^T$.;

7 The principal components of $\mathbf{X}$ can be computed directly from the principal components of $\mathbf{A}$: $\mathbf{W} = \mathbf{Q}\tilde{\mathbf{W}}$ and $\mathbf{\Phi} = \tilde{\mathbf{\Phi}}$.;

The projection of the data onto the principal component vectors $\mathbf{Y}_L = \mathbf{A}\mathbf{W}_L^\top$, the 'scores', are denoted by $\mathbf{Y}_L$ where the number of components maintained, L, is determined by the fraction of variance maintained. Columns from $\mathbf{Y}_L$ can be plotted as an image to examine trends extracted by PCA, and the spectral origin of these trends can be deduced from the principal component vectors.

3.2.2 Compressed PCA

Algorithm 3.2 (compressed PCA) was evaluated against computing PCA directly to demonstrate the equivalence of the results. Raw MALDI mass spectrometry images are too large to process directly so the fixed rat brain image introduced in Section 2.4 was spectrally rescaled at $\Delta m/z = 0.2$[66, 67], resulting in 4808 spectral channels. After this it could be loaded into memory and PCA computed using the `princomp` function in MATLAB. Note that this was only necessary for the purposes of comparison with standard PCA, the compressed PCA algorithm can process this dataset without re-binning. PCA was then performed using Algorithm 2.4 and Algorithm 3.2, for a range of values of k, which provide a range of compression ratios. The error metrics of SNR and PCC were calculated between the first five principal component coefficients generated directly and those generated using compressed PCA, five components were chosen as this corresponded to 95% of the variance in the data.

Computing the PCC and SNR differences between direct calculation of PCA and compressed PCA shows that the compressed algorithm successfully computed the principal components. Figure 3.1 shows how the

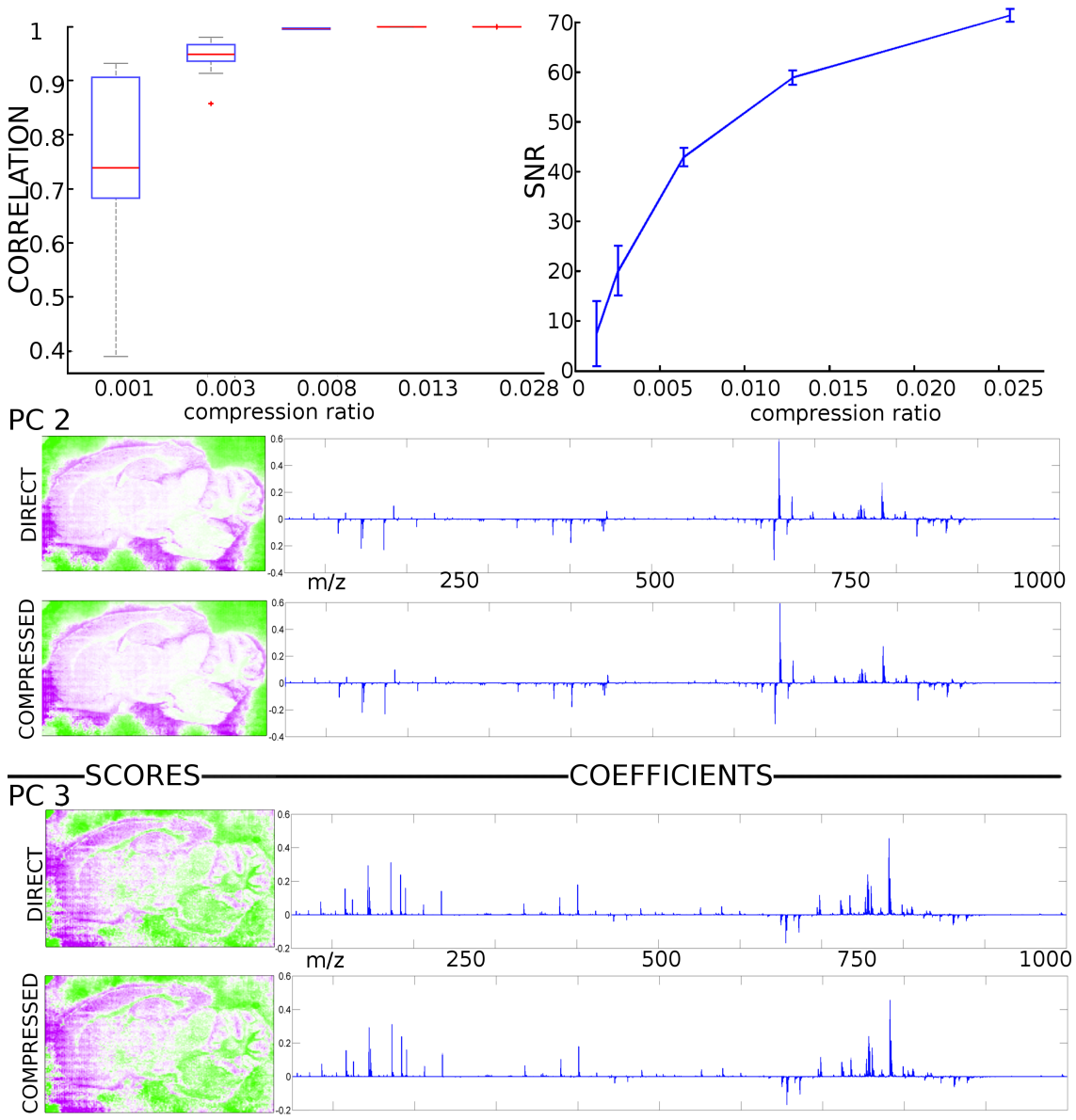

Figure 3.1: Comparison of direct and compressed PCA (compression ratio 0.0026) on a single dataset. Top: PCC and SNR error metrics (boxplot and error bars show variance from 10 repeats). Bottom: Scores and coefficients for principal components 2 & 3. High SNR and PCC coeffients of >0.99 show that the PCA coefficients generated on compressed data and decompressed are nearly identical to coefficients generated directly. This can be seen visibly by examining the scores and coefficients from PC 2 & 3 (PC1 was also identical).

SNR and PCC values vary with the compression ratio. Both the PCC and SNR metrics show an improvement as the compression ratio increases, but the PCC increases more rapidly and plateaus at a value greater than 0.99. Once the data is sufficiently highly sampled the variance from the median PCC tends to zero and the probability of isolated outliers is low. Below this value both the variance and the PCC value increase smoothly. The SNR continues to improve even after the PCC plateau is reached suggesting that thereafter any discrepancies are on the order of the noise. With the exception of the very lowest compression ratios chosen (corresponding to $k = 5$) the median value of the average PCC between the principal component coefficients obtained from the two methods over ten repeats was greater than 0.99. From this it is concluded that provided that a sufficient value of k has been chosen PCA performed on the compressed data gives the same results as PCA performed on the raw dataset within the noise limits of the data. Figure 3.1 also presents the coefficients and scores computed from PCA on the original data and compressed PCA using a compression ratio of 0.0026 illustrating that the interpretation of the coefficients can be performed in the m/z domain, as the coefficients have been back-projected.

It is also noted that the mean-centring of the data that is required can be performed whilst the data is compressed, $\bar{\mathbf{X}} = \mathbf{X} - \bar{\mathbf{x}}$ where the mean $\bar{\mathbf{x}}$ is subtracted column-wise, due to the choice of a symmetric distribution for the random matrix. Other schemes for approximate PCA [85, 237] stipulated that the original data must be mean centred prior to basis approximation, meaning that two more data passes are required: one to calculate the mean and once more to subtract it. Being able to do this on the compressed data then clearly provides a considerably improved pass efficiency. Variable standardisation, using sample z-scoring $\mathbf{Z} = \frac{\mathbf{X} - \bar{\mathbf{x}}}{\mathbf{s}}$ where $\mathbf{s}$ is the standard deviation, can also be performed on the compressed data as relative standard deviation is also preserved. Standardisation means that PCA is performed on correlation rather than covariance which may be useful due to the large magnitude variations between peaks seen in mass spectrometry data.

The time required for computation was also recorded during this comparison. Applying PCA to the uncompressed data took 1.5 hours on a desktop computer (MATLAB 2009a (32 bit), Mathworks), not including the time required to re-bin the data. Using the compressed method and a compression ratio of 0.0026, PCA took less than 8 seconds. This time includes the basis generation, compression of the data and back-projection of the eigenvector coefficients. This illustrates that even on datasets that can be processed directly the benefits of performing the component analysis on a compressed data matrix outweigh the additional overhead of performing the basis approximation.

Compressed factorisation for evaluating sample preparation

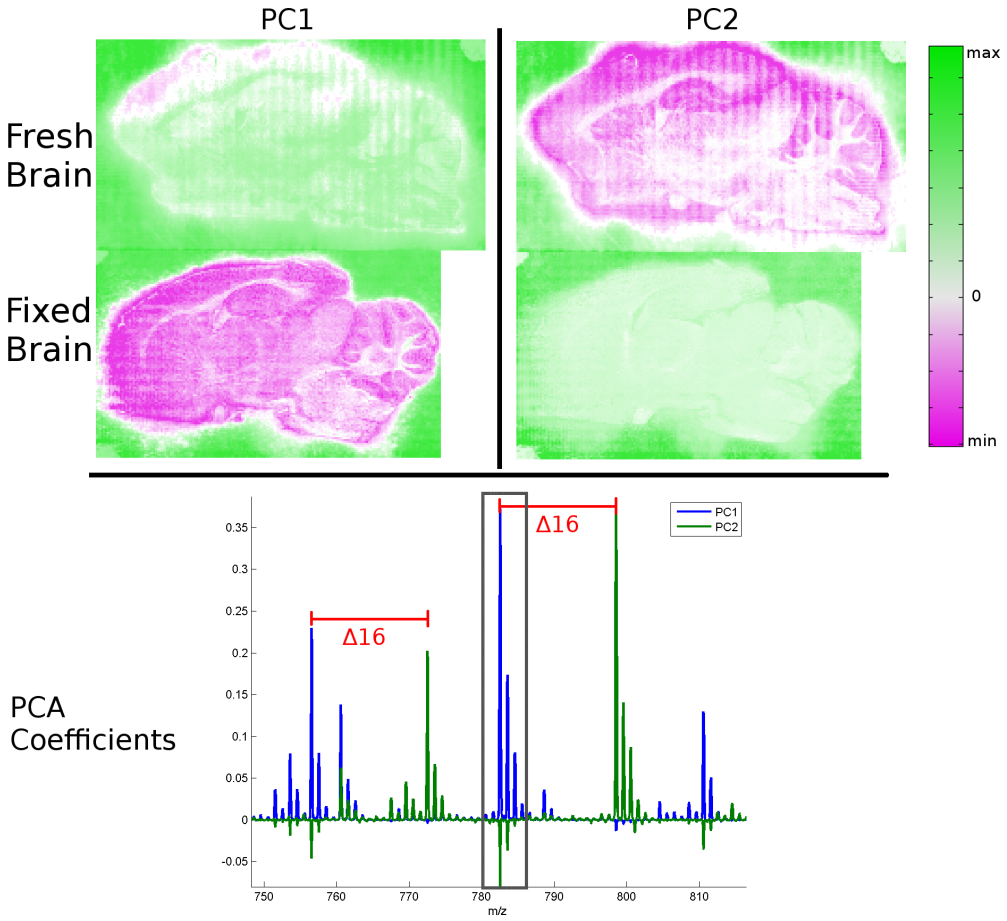

Figure 3.2: After combining two MALDI MSI datasets it was possible to perform PCA and examine the differences between the datasets. From principal components one and two it is clear that the two tissue types have been clearly highlighted as being different from each other, and their surroundings. Examining the coefficients gives a clearer idea of the source of the variance that PCA has identified: fixation shifted a selection of peaks by 16 mass units (as illustrated on the coefficient plot); this corresponds to a change in the lipid's adduct from potassium (39Da) to sodium (23Da). The decreasing series of peaks is characteristic of the isotope pattern of the lipids detected. This is explicit in the coefficients as the alternative adduct features negatively in the coefficient plot(highlighted with a box for m/z 782.5)

The merging of basis models introduced in Chapter 2 was used to combine the two rat brain images from Section 2.4. These are two mass spectrometry images which were collected separately. The datasets were acquired from rat brain processed in two fashions: one was fresh frozen rapidly following excision and the other was stored in a solution of neutral buffered formalin prior to freezing. The images were collected from sections revealing similar anatomy and so should be compositionally similar so any spectral differences are likely to be a result of the tissue handling. Compressed PCA was applied to the combined datasets and the images of the

scores from the first two principal components are plotted in Figure 3.2. Interestingly, the matrix background shows as being highly similar between the first two PCs suggesting that the matrix application was consistent between the samples. The fixed brain shows strongly in the positive component of PC1 whilst the fresh brain is dominant in PC2. This shows immediately that the sample handling has produced spectral differences between the two datasets. Examining the coefficient for these principal components, after decompressing which returns them to the m/z domain, stark differences were revealed. Several groups of peaks were seen to have been shifted by 16 mass units. This corresponds to the change in the metallic adduct detected in tandem with the lipid, potassium is seen predominantly with fresh tissue and sodium with fixed [32]. This is a rapid way of visualising the information content of the two datasets without requiring manual inspection of multiple ion images.

3.2.3 Non-Negative Matrix Factorisation

Non-negative matrix factorisation (NNMF) is used to decompose the data into a set of purely positive additive, coefficients and scores. This can be very useful for interpreting a system where each spectrum can be considered to be a mixture of a definite number of pure sources which cannot feature negative components[154]. This approach has been proposed for mass spectrometry imaging as the detected spectra have the physical property of being positive[108].

Recall from Chapter 1 that in NNMF the data, $\mathbf{X}$ is decomposed into the form (equation 1.9)

$$*\mathbf{X} \approx \mathbf{EG} \tag{3.1}$$

where $\mathbf{E}$ and $\mathbf{G}$ contain spectral coefficients and pixelwise scores respectively, both are constrained to be positive. As spectral compression using basis approximation results in negative values in the compressed data NNMF cannot be applied directly. It would be preferable to operate on the compressed data in a similar manner to PCA to reduce the computational burden and take advantage of the dimensionality reduction that has already been achieved. The semi-non negative matrix factorisation (s-NNMF) has been proposed as a solution for systems where the coefficients contain negative values but are still present as a mixture, removing the positivity constraint on $\mathbf{E}$ whilst maintaining it for $\mathbf{G}$. An algorithm for efficiently solving this problem has been presented [55] and is available online as a MATLAB toolbox `http://cs.uwindsor.ca/~li11112c/nmf.html` (v1.3).

Applying the s-NNMF decomposition to a compressed dataset, $\mathbf{A}$ yields the mixture fractions, $\mathbf{G}$ and

coefficients $\bar{\mathbf{E}}$ in the BASC basis domain:

$$\mathbf{A} \approx \bar{\mathbf{E}}\mathbf{G} \tag{3.2}$$

by pre-multiplying with $\mathbf{Q}$ and substituting in Equation 3.1 the coefficients can be recovered as in the original m/z domain.

$$\mathbf{QA} \approx \mathbf{Q}\bar{\mathbf{E}}\mathbf{G} \tag{3.3}$$

$$\mathbf{EG} = \mathbf{X} \approx \mathbf{QA} \approx \mathbf{Q}\bar{\mathbf{E}}\mathbf{G} \tag{3.4}$$

$$\mathbf{EG} \approx \mathbf{Q}\bar{\mathbf{E}}\mathbf{G} \tag{3.5}$$

$$\mathbf{E} \approx \mathbf{Q}\bar{\mathbf{E}} \tag{3.6}$$

Choosing the matrix rank

In contrast to PCA, NNMF is a factor analysis technique that requires a model of the data to be provided before the algorithm can be applied, including the number of factors present. Unfortunately, this is rarely available in advance, especially for a exploratory datasets. To estimate the data rank for the NNMF problem the eigenvalues generated by BASC-PCA were examined, looking at where the gradient of the eigenvalue curve asymptotes to zero (the scree plot). A value of 9 was selected from the scree plot and confirmed by running NNMF with 10 factors, which then produced a rank deficient output. There is not such a sensitivity to the number of random samplings used in basis approximation as an overestimate does not penalise the quality of the basis. Whereas, a poor estimate of the number of components can substantially degrade the quality of factor analysis methods[63]

Compressed s-NNMF was evaluated through a comparison to NNMF performed directly on the data (using algorithms from the nnmf toolbox version 1.3[131]). It was not possible to apply the NNMF algorithm to the whole fixed rat brain dataset, due to similar memory constraints that prevented the application of PCA directly. To provide a comparison set spectral rescaling at $\Delta m/z = 0.2$ was again performed. The compressed s-NNMF algorithm can be used on a full dataset without rescaling, but for this evaluation was also applied to the spectrally rebinned data. Computing NNMF on the reduced dataset (4751 channels) took over half an hour whilst performing compressed NNMF took less than three minutes, including compression and decompression.

Matching Scores for Comparison A disadvantage of NNMF is that whilst there is a unique best solution, the iterative solving algorithms used to search for it are very sensitive to the initial conditions and consequently

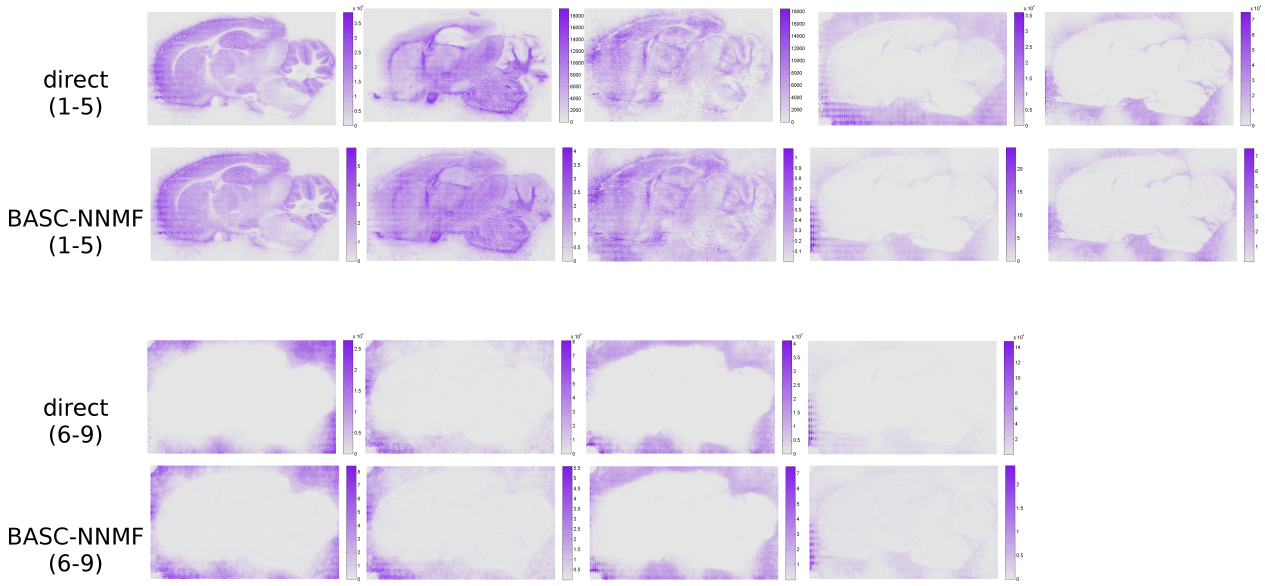

Figure 3.3: Comparison of the abundance maps produced by NNMF directly on raw data and compressed BASC. In both cases nine NNMF abundance maps were produced on data re-binned at $\Delta m/z = 0.2$ and then again folliwng BASC-compression. The colour on each map is scaled linearly between zero and the maximum value].

are prone to getting stuck in local minima[55]. This means that there is no way for the user to know whether a result produced is a global solution so repeated initialisations are required to expand the search space, keeping the best fit according to the smallest error metric $||\mathbf{A} - \bar{\mathbf{E}}\mathbf{G}||$. These replicates mean that additional time savings obtained by reducing the dimensionality of the problem are achieved for every replicate used. For all NNMF and compressed s-NNMF experiments in this chapter seven replicates of the algorithm were applied.

The initialisations are randomised so there is no defined order in which the maps from NNMF are produced. In order to compare results from different replicates and from compressed against full datasets the most similar abundance maps were matched using the following procedure.

Algorithm 3.3: match loadings across runs of factor loadings

Data: array of loadings, $\mathbf{G}$

Result: vector of matching indicies k

1 Calculate $\mathbf{P}$ the pairwise correlation coefficient between each pair of columns in $\mathbf{G}_1$ and $\mathbf{G}_2$;

2 Find the highest value in $\mathbf{P}$ and extract the row and column indices i, j;

3 set $k_i = j$;

4 Match the abundance map j to i and clear the column j and the row i (set the value to -1), repeat until all abundance maps are matched ;

The scores from direct-NNMF and compressed sNNMF with a compression ratio of 0.005 are shown side-by-side in Figure 3.3. Visual inspection of Figure 3.3 reveals that three factors corresponded to on-tissue distributions and the other four showed tissue-edge features and variation in the surrounding matrix. The positive values score maps can be interpreted as fractional abundances of the coefficients at each pixel. The scores from direct-NNMF and compressed s-NNMF with a compression ratio of 0.005 are shown side-by-side in Figure 3.3.

A systematic range of compression ratios (number of basis vectors) were trialled in the range of 0.003-0.05 (10-200 basis vectors). Each set of factors was matched to the most similar factor from the NNMF scores shown in Figure 3.3 using Algorithm 3.3. The correlation between the distributions was then calculated and is plotted as a function of number of samplings in Figure 3.4. There is a trend towards improving results as the number of projection increases (compression ratio decreases) which would be expected as the compression quality improves but there is a high level of variance at all points. One challenge in interpreting these results is that it is not known whether the factorisation on the full data is actually the best possible result that can be obtained. One observation that can be made is that the variance in the compressed s-NNMF results is more substantial than the compression errors seen from this dataset in Chapter 2. This suggests that variation from the initialisation conditions of the factorisation results tends to dominate and so it is essential to be able to repeat the factorisation many times to obtain an 'optimum' result.

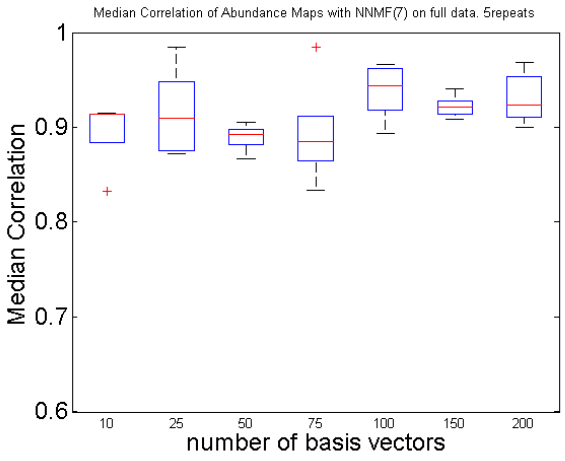

Figure 3.4: The correlation between NNMF abundance maps produced directly from the data and following compression. Typically a correlation of >0.9 is achieved regardless of the level of compression.

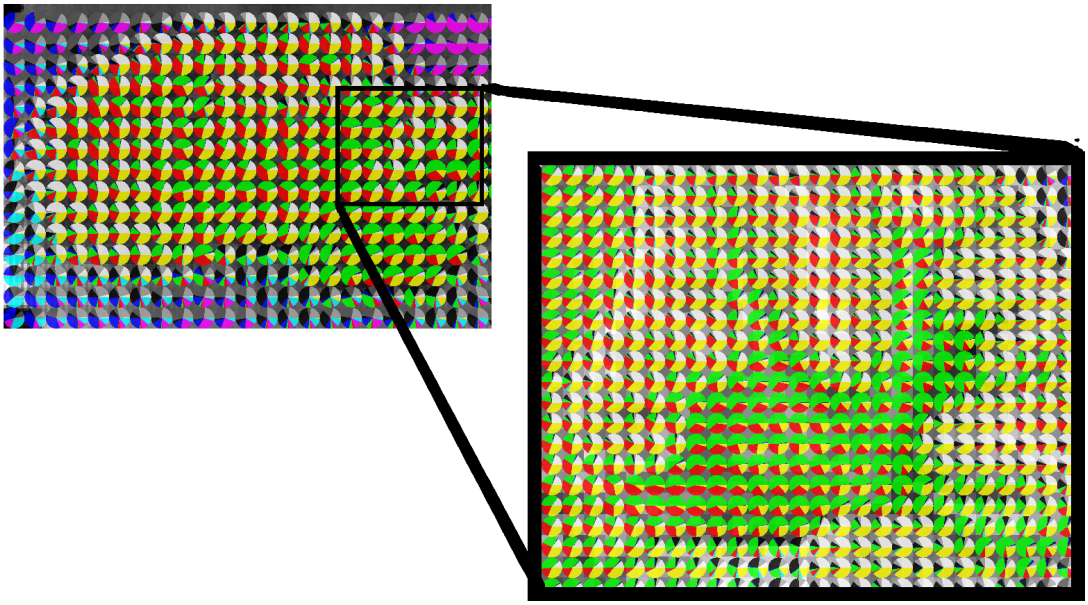

Figure 3.5: Simultaneous visualisation of all NNMF abundance maps. The segments of the circles represent the total fraction of the pixels under the circles belonging to each factor. By changing the circle radii, either dataset-wide visualisations can be produced or detail within regions can be seen.

Viewing the data

It is desirable to have a single overview of the factors, as it is difficult to visually assess abundance maps, especially as they may be on different colour scales for individual clarity. Recent visualisation work[67] required the embedding of the data onto three dimensions, so these could be directly transformed to red-green-blue channel intensities, which severely restricts the number of factors that can be introduced. The compressed s-NNMF approach (and factorisation methods in general) produce positive fractional abundances. This allows the production of single-view images, such as Figure 3.5, where the image region is overlaid with a grid of 'pie charts'. Each small circle uses a set of coloured wedges to show the fraction of that pixel attributed to each factor. By using a coarse grid on the whole image area an initial overview of the image is provided which clearly delineates the major features of the data, increasing the density of the map by zooming in or using a finer grid allows fine structure to be seen whilst preserving the knowledge from all of the multivariate components. In conclusion, this visualisation presents the quantitative elements of the multivariate analysis and provides the user with an overview of the structure of the data within a single image.

3.3 Compressed Segmentation

The estimation of distances between data points for use in segmentation algorithms is investigated. The size and complexity of mass spectrometry images, particularly from biological samples, has led to concerted efforts to provide user-friendly 'at a glance' information content visualisations. The most commonly used technique is segmentation, which divides an image into colour coded regions that are 'similar', sacrificing specific spectral information for spatial clarity. Visualisation approaches have been demonstrated for extracting subtle data features[67] but an advantage of segmentation is that a finite number of distinct image regions are produced, with characteristic spectra that are directly physically interpretable[2].

Segmentation divides the image up into areas of similarity, clustering provides an automated method for segmenting an image by grouping pixels with similar spectra together. Multiple similarity measures exist but the Euclidean distance is most commonly used and recommended for clustering[52, 54]. Pixels are usually considered independently during clustering so groups can be spatially discontinuous but the use of spatial pre-processing has been shown to improve clustering results [4].

Calculating the distances between the pixels in the image is the largest computational task for most clustering algorithms and performed naively, requires an element-by-element calculation against every measurement channel for every spectrum. As the Euclidean distance is preserved by the compression segmentation can be applied directly to the compressed data matrix ($\mathbf{A}_{m\times k}$) substantially reducing the number of calculations that need to be made without compromising the quality of segmentation. In fact, it may be expected that the quality could improve by calculating on the compressed representation as concentrating the measurements within the basis avoids the 'curse of dimensionality' and may suppress the influence of noise.

3.3.1 k-means Clustering

For details on the k-means algorithm see Section 1.4.5.

Segmentation with Compressed k-means Clustering

Segmentation of the MALDI rat brain image directly from the compressed data employing the popular k-means clustering[79, 108, 146], with 3-6 clusters being trialled was investigated. Algorithm 2.4 was applied for compression, using $k = 100$, then clustered using the MATLAB function `kmeans`.

The results of the segmentation using six clusters are shown in Figure 3.6. A segmentation map of the image area shows which cluster each pixel has been assigned to. A coloured map provides no information on how spectrally different the clusters are so the euclidean distance between every pair of cluster centroids are

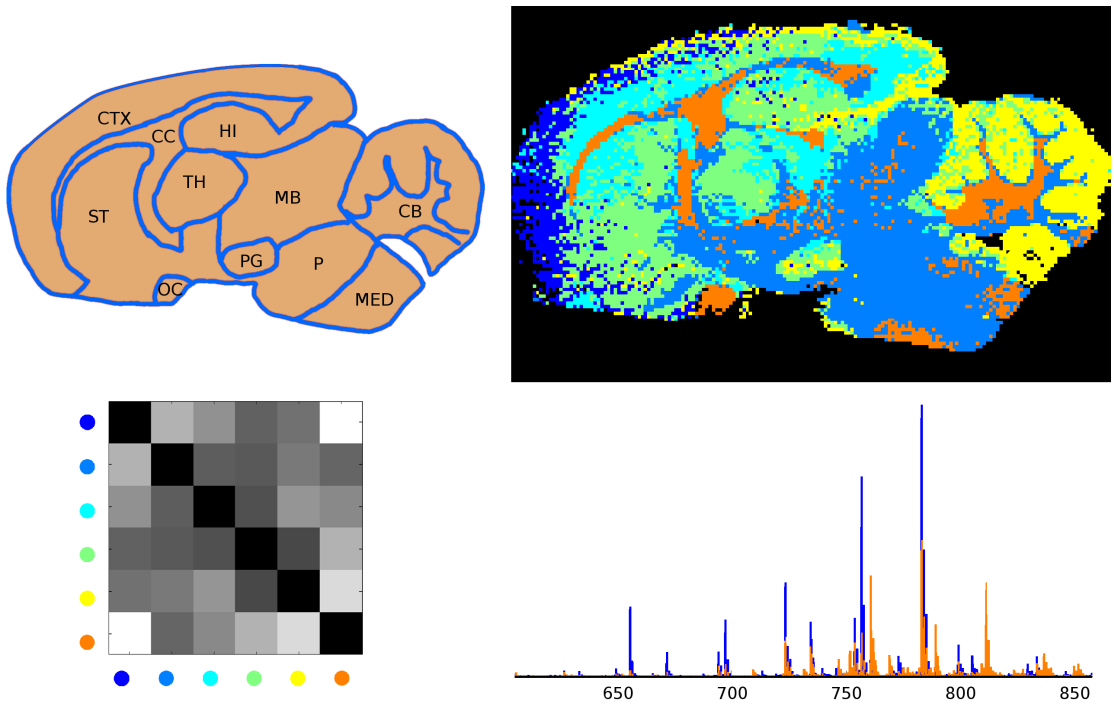

Figure 3.6: MALDI mass spectrometry image segmentation: k-means segmentation was used to segment the image into six regions and provide a simple map of spectral heterogeneity. Top left: Key histological structures (cerebellum (cb), cortex (ctx), hippocampus (hi), medulla (med), midbrain (mb), optic chiasm, pituitary gland (pg), pons (p), striatum (st), and thalamus (th)). Top right: segmentation map dividing the brain into six regions. Bottom left: Euclidean distance between the cluster centroids shown in the map (coloured accordingly). Bottom right: mean mass-spectra from clusters 1 & 6 (blue and orange respectively) showing clear differences in the peaks present from the lipid region.

shown on a grid. The tissue is divided along borders corresponding to known anatomical features, including the cerebellum in cluster 5 (yellow); the frontal cortex in cluster 1 (dark blue); the striatum in cluster 4 (green); and disconnected brain regions including white matter of the cerebellum, optic chasm and corpus collossum in cluster six (orange). The two most dissimilar clusters were identified from the grid as being clusters one and six. The mean spectrum for each cluster was generated and are shown overlaid, also in 3.6 allowing the spectral profiles of the clustered regions to be visualised.. The mean spectra are obtained by simply backprojecting the cluster centroids generated by the k-means algorithm.

Very similar regions are revealed as were seen when `kmeans` was applied following random projection (Figure 2.5), which is expected as they both preserve Euclidean distance as the similarity metric. The main difference is that the average spectra can be obtained directly from the clustering results whereas from the random projection result another pass through the data would have been required to calculate them. The workflow is seen to provide a rapid route to visual inspection of data distributions and spectral profiles from important image regions.

3.3.2 Self Organising Map (Self Organising Map (SOM))

The training is an iterative competitive reinforcement process and proceeds as described in the following algorithm(adapted from [119]):

Algorithm 3.4: Train the nodes in a SOM $\mathbf{N}_{s_x \times s_y \times k}$ so that they span the variation contained within a compressed dataset $\mathbf{A}_{m \times k}$

Data: node grid dimensions: s_x, s_y, compressed dataset projections $\mathbf{A}$

Result: node grid $\mathbf{N}$

1 Randomly initialise the weightings of every node on the grid e.g. pick values from a normal distribution $\mathbf{N} = N(1, 0)$.;

2 Pick (at random) a vector from $\mathbf{A}$: $\mathbf{a}_i$;

3 Find the most similar node, i.e. for $\mathbf{M}_{s_x \times s_y} = |\mathbf{N} - \mathbf{a}_i|$ find the indices s_a, s_b of the smallest value of $\mathbf{M}$;

4 Adjust the weighting of the $N(s_a, s_b) = N(s_a, s_b) - \alpha(\mathbf{N}(s_a, s_b) - \mathbf{a}_i)$, α is a factor controlling the strength of the change and is set to decrease as the training process continues;

5 Adjust the weighting of every node in the grid to make it more similar to the datapoint: $\mathbf{N}(i, j) = \mathbf{N}(i, j) - \alpha\beta(\mathbf{N}(s_a, s_b) - \mathbf{a}_i)$, β is a function that measures the node distance to s_a, s_b (in grid coordinates) $\beta(i, j) = 1/\left((a - i)^2 + (b - j)^2\right)$;

6 Repeat steps 2-3 decreasing the training strength for a fixed number of iterations;

Segmentation using a compressed SOM

Once the map is trained on the input data, each data point is classified by allocating them to the most similar node as measured by Euclidean distance. Several features of the SOM are worth noting: Only the projection matrix is required for the construction of the map, the resulting node weightings are expressed in terms of the projections and can be decompressed to the m/z by multiplying with $\mathbf{Q}$ for further interpretation. As the training process effectively aims to have an equal number of data points per node, the nodes end up spaced according to data density, which provides a higher colour dynamic range over dense data regions. Empty nodes are allowed to occur during classification even though the SOM is spaced over the data-space so not every node is necessarily represented in the resulting segmentation map.

Only the data-space required is encompassed by the SOM, which makes comparing separate SOM segmentation difficult without re-calculating the map. It is possible to classify new data by allocating it to the node with the closest weighting but datapoints will *always* be allocated a node even if it is very dissimilar spectrally. For comparison, the datasets should be combined (e.g. using basis merging, Algorithm 2.6) before the SOM is trained.

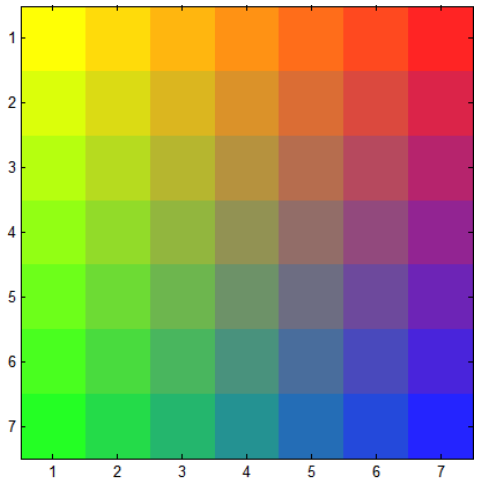

Figure 3.7: SOM colourmap for a 7×7 SOM node grid using a primary RGB colourspace. The colourmap is applied independently of the node weightings

Segmentation Colour Scheme

Visualisation of the segmentation is achieved by applying a colour scale onto the SOM nodes and transferring the colour values to the classified data. As the nodes are weighted so that nearby nodes have similar spectral weightings the colour scheme chosen should be smooth with nearby nodes coloured similarly. Figure 3.7 show the RGB colourmap used, where the corners were allocated the primary colours (plus yellow) and the intermediate nodes coloured by linearly interpolation.

As the colour-map strongly effects the resulting visualisation care should be taken in interpretation of the segmentation. Human colour perception is non-uniform over the chromatic scale[191] and as such colour normalisation may be required[80] to provide an un-biased discrimination whilst viewing. The non-linear density of node-spacing means that there is a non-linear difference as colour-shade changes. Some authors colour the nodes according to data values[136, 163] but this is only practical if the number of dimensions is equal to the SOM dimension, in this case where the dimensionality of the projected data is still >100 there is no obvious mapping and this would defeat the density adaptivity of the SOM.

Comparison of BASC-SOM with full-spectrum SOM

The self-organising-map is trained spectrum-by-spectrum in a memory efficient manner so it can be applied to a complete MSI dataset, allowing a side-by-side comparison of the results and performance of compressed analysis. This is possible as the SOM is trained sequentially on individual spectra so it can be implemented in a memory efficient fashion reading a single spectrum at a time from disk. For completeness, the rebinned

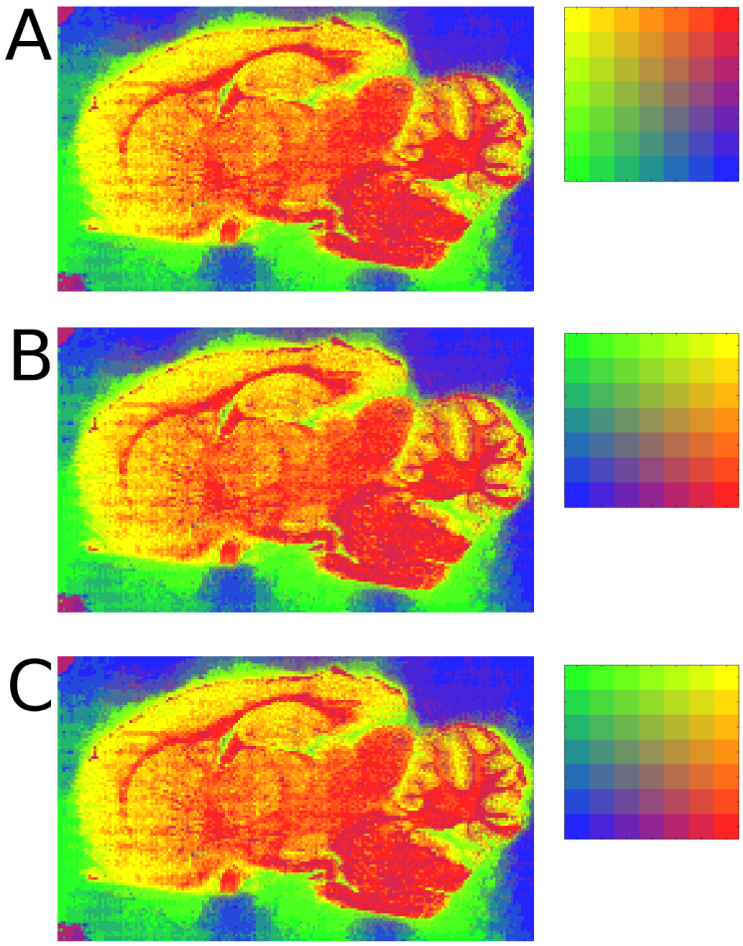

Figure 3.8: A. Example of SOM run on full data (stored as imzML), B. rebinned at $\Delta m/z$= 0.2, C. compressed (with 150 samplings) using BASC. Whilst generally similar results are obtained and tissue anatomy is identified in both cases the clustering patterns from the compressed data are much less noisy. The SOM node colours shown in the right-hand column, note the rotation of the colours applied to the first SOM in order to make the segmentation map colours match.

data was also included in the comparison.

Self organising maps (7×7 nodes) were trained independently on the full data, the same dataset after rebinning at $\Delta m/z = 0.2$ and on the BASC compressed data (using k=150) and the subsequent segmentation maps are shown in Figure 3.8. In normal operation, a SOM is randomly initialised and random spectra are chosen for training, to allow comparison the start point was fixed (and transformed between data representations as required) and the computer's random number stream was reset before each SOM was produced so that the same training points were chosen. As expected it was very slow to train the SOM on the full spectrum data and the whole process (including classification) took 9 hours. Operating on datasets stored in memory was much faster where it took 15 minutes for the reduced data and <10 seconds on the BASC projections. These timings do not take into account the time required to re-bin or compress the data, which are approximately 1 hour in each case, but there is still a substantial time saving. The SOM algorithm is inherently more pass efficient than that used in k-means, with the number of iterations and data-points per iteration specified, each data-point is used approximately twice on average during the training of the SOM, with one further pass for classification. By contrast,the k-means algorithm uses every data point once per iteration making it inefficient for data stored out of core memory. Even with this improvement, the SOM is still slow when accessing data from disk and it is well worth the time taken to perform basis approximation prior to segmentation.

The distributions visible in the SOM segmentation maps in Figure 3.8 reflect the tissue compartments visible from k-means segmentation and correspond to the known histology. A far greater number of SOM nodes were used than clustering centroids are typically allowed in other segmentation approaches, typically 5-10 depending on the number of tissue types expected[6, 54], which allows smaller features to be clustered separately and gradients between regions to be seen. More subtle image features are indeed visible in Figure 3.8, particularly within the cerebellum. These results are in agreement with the visualisation methods applied to this dataset by Fonville et al[67].

Once the number of measurements is lower than that number of samples, as was the case with the rebinned data, then distance metrics should be stable and the results following BASC would be expected to be similar. As the same spectra were used to train the map, corresponding nodes between the two SOMs share the same spectral profiles. As expected, Figure 3.8b-c shows that almost identical spatial patterns were extracted in each case. There is not quite such a similarity between the full data (Figure 3.8a) and the dimensionality reduced cases (Figure 3.8b-c). Note that the colour map applied to the SOM (the left column in Figure 3.8) has been manually adjusted to rotate the colour-scheme by 90 degrees. This then restores agreement between

the colours in the segmentation maps. It seems likely that the curse of dimensionality is active here and that some discrepancy in the distances calculated causes different nodes to be determined to be closest during training. Once the colour map was rotated similar regions were all coloured identically but the segmentation map is visibly more noisy than in the other two cases.

3.4 Conclusion

Basis approximation has been shown within this Chapter to provide more than simple data compression. It has been shown to augment feature extraction using factorisation and potentially provides an alternative as an input to segmentation algorithms. Factorisation is frequently used as a precursor to segmentation[67] so benefits remain even if segmentation of BASC directly is not performed. Other clustering methods using the Euclidean distance such as hierarchical clustering[2, 52] could also potentially benefit from random matrix-based compression, and future work will include a thorough investigation of this, and other methods.

Visualising the output of factorisation still remains a challenge for human perception, as multiple patterns must be simultaneously interpreted, the use of single view outputs provides an alternative method. Extending this data-view to factorisations with negative values, such as PCA, presents an interesting challenge. Segmentation provides a single view of data heterogeneity, of the methods investigated here, SOM presents a more sophisticated view of the data, at the expense of a greater number of clusters being produced. It can be applied inherently faster as the training process shows individual spectra to the nodes during training whilst k-means uses every data point at each iteration.

Factorisation and segmentation are often performed on feature extracted data such as the factorisations enabled using basis approximation allowing an unbiased examination of the data. As an approach that can be applied to the data without any prior processing BASC provides a baseline against which the effects of further processing can be compared.

Chapter 4

Comparison and Evaluation of Feature Selection Methods

This chapter explores the impact of feature selection on the factorisation and segmentation of mass spectrometry images. Literature-based pipelines are evaluated, both those which try and retain the majority of spectral information (re-binning and peak selection methods) and those which actively discard non-informative channels. Data processed directly with Basis Approximation for Spectral Compression (BASC) was used as a baseline for comparison. Evaluation of the effect of dimensionality reduction was made with sparsity measure and data visualisation using a Self Organising Map (SOM).

Pipelines that aimed to reduce the data size whilst capturing the majority of spectral information were determined to be largely successful in reproducing the tissue features visible in the data compressed with BASC. However, methods that minimised processing time by using a subset of the data were not able to reproduce as much tissue heterogeneity during visualisation.

4.1 Introduction

An important stage in automated processing pipelines is feature selection and extraction where significant measurements are determined and a useful subset chosen. Feature selection is the process of detecting informative measurements and feature extraction is the process of choosing or combining informative measurements for further analysis. A great number of algorithms and pipelines for feature selection and extraction of Mass Spectrometry Imaging (MSI) data have been developed and presented (e.g. [4, 54, 67, 148]). Information can be lost at this point so it is important to understand the effect that feature selection and extraction has on subsequent processing and establish methods for evaluating the available algorithms. The key question is how to evaluate the effect that feature selection and extraction has on downstream data analysis, and what criteria to use for selecting a pipeline.

Studies have compared the similarity of the outputs from feature extraction methods, including factorisations (Non-negative matrix factorisation (NNMF), Principal Component Analysis (PCA) and pLSA) and segmentation (k-means and fuzzy k-means)[67, 108]. In both cases, a feature selection stage in the processing pipeline was used (peak detection in [108] and spectral re-binning in [67]) but no evaluation of the effect of the feature selection stage was undertaken and they did compare against a baseline of no processing. As these algorithms can only operate on the features selected it is important to understand how much influence the feature detection and selection algorithms have. The two main factors hindering inter-laboratory evaluation are, firstly, a lack of fully featured extensible software has left individual labs to develop their own solutions that are non-trivial to share[2] and secondly there are few publicly available datasets which allow direct comparison of methods from multiple groups. By establishing methods for comparing algorithms inter-laboratory studies, which are not commonly performed, could be made possible.

This work explores several metrics for evaluating the impact of feature selection methods in MSI:

1. The running characteristics can effect the choice of which algorithm to use so any comparison and evaluation of algorithms should include consideration of their relative time and computational behaviour. Some factors to consider include the total run time, pass efficiency and memory efficiency. The main quantitative comparisons that can be made between selection/extraction routes are, the number of measurements to consider; the spatial and spectral sparsity of the measurement set; the variance contained in the principal components and the time requirements of each pipeline.
2. A 'molecular histology' approach extracts spectral patterns based on the delineation of tissue regions which can be automatically determined using (amongst other approaches) spatial segmentation [4]. As segmentation is based on automatic detection of spectrally similar areas it provides as useful way of

seeing the effect of different feature selection pipelines. It is important to establish whether any pipeline alters the tissue compartments detected.

The compressed analysis introduced in Chapters 2 & 3 can be applied directly to data that has not undergone any feature selection and so provides a baseline against which alternative processing can be evaluated. This work presents a route to producing recommendations for the relative suitability of pipeline as well as exploring specific strengths and weaknesses of existing pipelines.

4.2 Constituent Algorithms

Data processing pipelines combine several algorithms into a single workflow, the complete pipelines are then described in Section 4.3. As some algorithms are included in more that one pipeline they are detailed here individually.

4.2.1 Pre-Processing

A standard pre-processing stage was defined and applied to each spectrum independently to de-noise and restore the data to a continuous spectral axis. Spectra were stored in imzML format as discontinuous mass to charge ratio (m/z)/intensity pairs and loaded using the parser from imzMlConverter[173].

Algorithm 4.1: Preprocessing stages

Data: a vector of m/z values **m** and corresponding intensity values **c**, global mass axis **g**

Result: preprocssed spectrum intensity values **x**

zero-filling

1 Initialise **x** as a vector of zeros of the same size as **g**;

2 Calculate a vector **v** of indices for the values of **m** within **g**;

3 Transfer the values of **c** to **x** at the indices **v**;

denoising

4 SavitzkyGolay filter applied to **x**;

Any empty values in the continuous intensity vector were zero-filled. A small amount of de-noising was applied using a second order SavitzkyGolay filter[184], with a fixed width of 25 m/z bins. The filter width was fixed at 25 m/z bins which was manually determined to be approximately the width of a well defined peak at $m/z \approx 650$.

4.2.2 Normalisation

Normalisation attempts to compensate for inter-spectra variability that is unrelated to the underlying signal. Total Ion Chromatogram (TIC) normalisation assumes that approximately the same number of ions should be detected in each spectra

Algorithm 4.2: TIC normalisation

Data: input dataset $\mathbf{X}$

Result: normalised dataset $\bar{\mathbf{X}}$

for *j=1* **to** *number of columns in* $\mathbf{X}$ **do**

  divide each spectrum within $\mathbf{X}$ by its sum: $\bar{\mathbf{X}}_j = \frac{\mathbf{x}_j}{\sum \mathbf{x}_j}$;

end

4.2.3 Re-binning

Re-binning is typically used to shorten the m/z axis by re-sampling the rate at which m/z measurements are taken. This can acheive some de-noising through averaging[212] and also achieves dimensionality reduction. These undirected feature selection methods return a user-defined number of peaks (defined by the new bin width) which needs to be chosen with careful inspection of the data so that a bin width is not chosen that begins to merge peaks.

Algorithm 4.3: Data re-binning

Data: original global mass axis $\mathbf{g}$ and corresponding vector of intensity values $\mathbf{x}$, new mass axis $\mathbf{m}$

Result: rebinned vector of intensity values $\mathbf{y}$

initialise $\mathbf{y}$ vector of zeros of the same size as $\mathbf{m}$

for *j=1* **to** *length of* $\mathbf{g}$ **do**

  determine the index, i, of the bin in $\mathbf{m}$ axis which $\mathbf{g}_j$ falls within;

  increase $\mathbf{y}_i$ by $\mathbf{x}_j$;

end

A linearly increasing m/z axis with $\Delta m/z = 0.2$ was produced between the minimum and maximum values in the original data and Algorithm 4.3 applied. Re-binning was performed on each spectrum independently using the same axis for every spectrum. When possible within computational constraints the results were stored in memory as an array.

After linear re-binning it is no longer possible to return to the time domain, where the intensities were originally recorded on a linear measurement axis which may complicate further spectral processing that relies on an even spacing.

4.2.4 Peak-Picking

Peak picking determines centroid values of peaks within the data. All signal under a particular peak is then assumed to be produced by the detection of a single ion species[230]. Peak apexes were identified by determining intensity values within $\mathbf{x}$ which were the larger than every other value within a neighbourhood window[233]. The window was 7 m/z bins wide, which was the Half Width Half Maximum (HWHM) of a well defined peak at $m/z \approx 650$.

4.2.5 Summary Spectra

Summary spectra is a term used for descriptors of m/z channels where the descriptor is compiled for each channel independently over the whole dataset. An implementation of each algorithm was found from the literature which allows the calculation to be computed in a single pass over the data[42, 148, 224]. Calculation of statistical dataset summary spectra was introduced by Coombes et al[42] and extensions with a specific focus on imaging data were presented by McDonnel et al[148]. The variance of each spectral channel is visually displayed in some commercial software (SCiLS Lab 2014a).

Standard Deviation Spectrum The standard deviation of every m/z channel can be calculated with a single pass through the data using the on-line variance algorithm as in [224]:

Algorithm 4.4: Single pass standard deviation spectrum

Data: mass spectrometry image $\mathbf{X}$

Result: standard deviation spectrum $\mathbf{s}$

initialisation

$n = 0$;

$\mathbf{m}$ = vector of zeros equal to the number of columns in $\mathbf{X}$ (number of m/z channels);

$\mathbf{v}$ = vector of zeros equal to the number of columns in $\mathbf{X}$ (number of m/z channels);

online variance algorithm

1 **for** $i = 1$ **to** *number of spectra* **do**

2 $\quad \mathbf{x}_i$ = read ith spectrum;

$\quad$ update variance estimate $n_i = n_{i-1} + 1$;

4 $\quad \mathbf{d} = \mathbf{x}_i - \mathbf{m}_{i-1}$;

5 $\quad \mathbf{m}_i = \mathbf{m}_{i-1} - \mathbf{d}/n$;

6 $\quad \mathbf{v}_i = \mathbf{v}_{i-1} + \mathbf{d}(\mathbf{x}_i - \mathbf{m}_i)$;

end

calculate standard deviation $\mathbf{s} = \sqrt{\frac{\mathbf{v}}{n-1}}$;

Mean Spectrum

Algorithm 4.5: Single pass mean spectrum

Data: mass spectrometry image $\mathbf{X}$

Result: mean spectrum $\mathbf{s}$

initialisation $\tilde{\mathbf{x}}$ = vector of zeros equal to the number of columns in $\mathbf{X}$ (number of m/z channels);

1 **for** $i = 1$ **to** *number of spectra* **do**

2 $\quad \mathbf{x}_i$ = read ith spectrum;

3 $\quad \tilde{\mathbf{x}} = \tilde{\mathbf{x}} + \mathbf{x}$;

end

calculate mean $\tilde{\mathbf{x}} = \frac{\tilde{\mathbf{x}}}{i}$;

Base-peak Spectrum

Algorithm 4.6: Single pass basepeak spectrum

Data: mass spectrometry image $\mathbf{X}$

Result: basepeak spectrum $\mathbf{s}$

initialisation $\hat{\mathbf{x}}$ = vector of zeros equal to the number of columns in $\mathbf{X}$ (number of m/z channels);

for $i = 1$ **to** *number of spectra* **do**

  $\mathbf{x}_i$ = read ith spectrum;

  compare each m/z channel against the maximum previously seen $\hat{\mathbf{x}} = max([\hat{\mathbf{x}}, \mathbf{x}])$;

end

4.2.6 Centroid List Merging

When peaks lists are produce from multiple spectra they must be combined into a single dataset list. As the same peak can appear in more than one list, with some experimental variation on the exact centroid m/z these lists must be combined with some opportunity to merge entries.

Algorithm 4.7: Single pass datacube construction

Data: lists of peak centroids $\mathbf{p}_1, \mathbf{p}_2 ... \mathbf{p}_n$, filter width f

Result: aligned centroid list $\mathbf{a}$, frequency count for each centroid

concatenate all $\mathbf{p}$ vectors into a single list, $\mathbf{a}$ and sort in ascending order;

calculate the difference between adjacent values, $\Delta\mathbf{a}$;

while $min(\Delta\mathbf{a}) < f/4$ **do**

  find i, the index of smallest value of $\Delta\mathbf{a}$;

  replace $\mathbf{a}_i$ with the average the value of $\mathbf{a}_i$ and $\mathbf{a}_{i+1}$;

  replace the frequency count of $\mathbf{a}_i$ with the sum of $\mathbf{a}_i$ and $\mathbf{a}_{i+1}$;

  recalculate $\Delta\mathbf{a}$;

end

The m/z tolerance value, f, is a peak width estimated from the data and is defined in terms of the base width as used in data-cube construction (Algorithm 4.8).

4.2.7 DataCube Construction

Datacubes are a set of ion images produced over a whole dataset. They are defined by a peak centroid list and a filter width, which corresponds to the width of a peak. In this way a high-dimensional dataset is

reduced to a small number of features.

Algorithm 4.8: Single pass datacube construction

Data: mass spectrometry image $\mathbf{X}$, global mass axis $\mathbf{g}$, peak list $\mathbf{p}$, filter width f

Result: datacube $\mathbf{Y}_{p \times n}$

for $i = 1$ **to** *number of spectra* **do**
  $\mathbf{x}_i$ = read ith spectrum;
  for $j = 1$ **to** *length of* $\mathbf{p}$ **do**
    $\mathbf{Y}_{ij} = \sum_{\mathbf{g}>\mathbf{p}_j-f}^{\mathbf{g}<\mathbf{p}_j+f} \mathbf{x}_i$;
  end
end

The filter width was heuristically determined to be $f = 0.05m/z$ based on the width of a peak at $m/z \approx 650$.

4.3 Pipelines

Each pipeline consists of a series of algorithms in a specific order, this section describes the ordering of several pipelines adapted from the literature into a single computing environment (MATLAB 2013a). The steps of the constituent algorithms are described separately in Section 4.2. In each case a final stage of feature extraction will be performed using BASC-PCA and a SOM will be used to visualise the results. Some considerations regarding implementation are discussed along with the description of the pipelines. In particular, whether the algorithm can be deployed in a memory and computationally efficient fashion. These must be evaluated at a pipeline level as it is possible to combine some operations within a single spectrum access event in order to minimise the time spent reading data. As implementation was performed in MATLAB additional absolute time savings may be achieved with a full compiled programming language, for this reason relative characteristics are compared.

In these descriptions the pipelines are numbered by operations that require a data pass, or an offline calculation.

4.3.1 Pipeline: None (BASC)

The data was compressed without further manipulation using the BASC algorithm as described in Chapter 3 (Algorithms 2.4). The data compression rate was chosen based on work in earlier chapters on this dataset (k=200).

1. Pre-processing (Algorithm 4.1)
   Basis Construction (Algorithm 2.4)

2. Projection onto BASC basis

Efficiency

This method is memory efficient by design and also provides a pass efficient approach, where the pre-processing was applied on-line to each spectrum as it was loaded into memory.Following basis generation and compression no further data access is required. Additionally, as described in Chapter 3, the data objects that make up the compressed form are used independently for computation of PCA so do no even need to be stored in memory simultaneously.

4.3.2 Pipeline: Rebin (BASC)

Data was rebinned along the spectral axis into m/z bins of fixed width $\Delta m/z = 0.2$, consistent with common commercial software conversion[67]. To implement this in a memory efficient manner, during a first pass of the data each spectrum was loaded independently and re-binned whilst being used to generate the approximate basis. A second pass (also with on-line rebinning) was used to project the data. This approach was found to best respect memory limits in limited computational environments.

1. Pre-processing (Algorithm 4.1)
   Re-binning (Algorithm 4.3)
   Basis Construction (Algorithm 2.4)

2. Projection onto BASC basis

Efficiency

This implementation uses the basis approximation framework for the purposes of pass efficiency, so re-binning was performed spectrum-by-spectrum during basis generation and again during projection. This gives it the same efficiency characteristics as the raw data. As the rebinning can be applied to each spectrum individually it was possible to perform the basis approximation sampling whilst the spectrum was still in memory giving this method the highest overall pass efficiency.

4.3.3 Pipeline: Standard deviation spectrum

Producing a single summary spectrum for use with peak detection is a commonly used method for producing a high Signal to Noise Ratio (SNR) spectrum and avoiding spectrum-by-spectrum alignment issues[157], here the 'standard deviation spectrum' was produced.

Peak picking was then performed on the standard deviation spectrum to produce a list of detectable peaks and a datacube produced from the resulting m/z list.

1. Pre-processing (Algorithm 4.1)

   Standard deviation spectrum construction (Algorithm 4.4)

2. Peak detection on summary spectrum (Section 4.2.4)

3. Datacube construction (Algorithm 4.8)

Efficiency

A pass is required for producing the standard deviation spectrum and another for producing the data-cube. No restrictions were placed on the number of peaks that could be detected. Assuming a data-cube could be stored in fast memory it must be accessed three times to perform PCA (once to produce a mean spectrum, once for eingenvalue decomposition and finally for producing the scores). As the BASC algorithm operates on a spectrum-by-spectrum basis it could still be used for dimensionality reduction in the case where so many peaks are detected that datacubes cannot be stored in memory.

4.3.4 Pipeline: Multiple summary spectra

Peak detection on multiple summary spectra allows higher sensitivity to peaks that span a range of magnitudes within the data, compared to peak picking on the mean spectrum alone[147]. The mean spectrum, skyline (or base-peak) spectrum and their TIC normalised versions (as described in [147]) were produced and peak picking was performed on each individually. The resulting mass lists were pooled and a datacube was then formed from this list.

1. Pre-processing (Algorithm 4.1)

   Mean Spectrum construction (Algorithm 4.5)

   Base-peak (or skyline) spectrum construction (Algorithm 4.4)

   TIC-Normalised Mean spectrum construction (Algorithm 4.5)

   TIC-Normalised Base-peak spectrum construction (Algorithm 4.4)

2. Peak detection on summary spectrum
   Merging of peak lists (Algorithm 4.7)

3. Datacube construction (Algorithm 4.8)

The only parameter required for this method is the peak width $f = 0.05$ that was established as the width of a well defined peak at $m/z \approx 650$

Efficiency

All of the summary spectra were produced during a single pass of the data maintaining only one spectrum at a time in memory. The non-normalised summary spectra were updated first, the spectrum was then overwritten element-by-element with its normalised version and the normalised summary spectra updated. A second pass was then required to construct the data-cube. As with all data-cube based methods, three further passes of the cube were then required for factorisation with PCA.

4.3.5 Pipeline: Frequent peaks

Peak picking on every spectrum is a computationally intensive task[233] and so to reduce the burden it has been suggested that taking a subset of the image provides a compromise between coverage and speed[4]. From each peak-picked spectrum the most intense peaks are recorded and their frequency of appearance tallied over the dataset. Peaks that appear often are assumed to be most informative and so maintained to form the final peak list.

1. On subset of spectra: pre-processing (Algorithm 4.1)
   Peak detection

2. Merging of peak list (Algorithm 4.7)
   Discard peaks that appear infrequently

3. Datacube construction (Algorithm 4.8)

Parameters were taken from [5] and so the percentage of pixels to peak-pick was 15%, and spaced evenly over the image; the top 20 peaks from each spectra were tallied; and peaks contained within 1% of spectra were kept. For peak merging $f = 0.05$ was again used.

Efficiency

To build a peak list 15% of the spectra were taken using a uniform sampling over the image and a list of peak centroids recorded for each. This allows the peak list to be generated without a full parse of the dataset. As the number of peaks per spectrum is low (20) the total m/z list never grows too large meaning there is a high likelihood of the resulting datacube fitting into memory. A single full data pass was then required to build the data-cube.

4.3.6 Pipeline: Spatial Correlation

The general method proposed in[66] is a two-stage peak filtering scheme that first removes m/z bins that correlate with designated matrix peaks and then eliminates any channels that show low internal variance.

1. Pre-processing (Algorithm 4.1)
   Rebinning (Algorithm 4.3)
2. User manually reviews the re-binned data and selects a number of 'background' or 'non-informative' ion images
3. The Pearson's Correlation Coefficient (PCC) is calculated between every channel and the list of background ion images
   Channels with high average correlation are discarded
4. PCA of every ion channel (image rows as measurements and columns as samples)
   Low variance channels are discarded

The original paper[66] suggested linear re-binning at 0.2 m/z. Other parameters that are required are a threshold for variance to keep and the number of matrix peaks to correlate against, these were left as the default 20 and 10 respectively.

Efficiency

Spatial correlation requires the first data pass to re-bin the image sufficiently so that it can fit into memory. As such this was the only method that was not implemented in a memory-efficient manner. This makes it pass efficient in the case that the data can be reduced sufficiently but otherwise forming each ion channel for correlating requires a pass though the data making this method much less attractive.

4.4 Pipelines Compared on Real Data

Name	Memory Efficient	Disk Passes	Memory Passes	Time (relative 'none')†
None	Y	2	0	1
Re-Binning 0.2	Y	2	0	0.14
Standard Deviation Spectrum	Y	2	3	0.20
Multiple Summary Spectra	Y	2	3	0.22
Frequent Peaks	Y	1.15	3	0.15
Spatial Correlation	N	1*	3	0.09

Table 4.1: Pass efficiency analysis of the pipelines plus dimensionality reduction, as implemented within this work. Numbers shown for processing starting at raw data. † timings are approximate and no code optimisation was attempted. * a first stage of feature selection (re-binning at $0.2m/z$) was specified for this method so that the data small enough to fit into memory, if this cannot be achieved then this method becomes substantially less pass-efficient.

This section will examine the application of each of these pipelines to a real-world MALDI image. It will compare the output from the feature extraction in terms of number and nature of features retained and the final 'molecular histology' results obtained by segmentation. There is not a particular criteria for the number of measurements retained, but for effective calculation of distances a rule of thumb is that the sample to measurement ratio of 10 is not exceeded [195], comments from other authors suggest that 100-200 measurements should be retained[2]. More important is that the measurements kept are discriminatory between tissue regions so that histological differences are revealed. Using a real Matrix Assisted Laser Desorption Ionisation (MALDI) dataset comes with drawbacks, other than the data size, most notably the lack of a ground truth against which to compare any results which leads to a narrative conclusion. Fortunately as BASC can be applied to the data without prior reduction, a comparison can now be made against a baseline of no processing giving an absolute analysis of the effect of each pipeline. Several types of efficiency can be considered but the largest hurdle for mass spectrometry for day-to-day use the total time is the largest concern[2]. As a major bottle-neck for imaging data can be disk load time, the pass efficiency (number of times a dataset must be read) can have a substantial impact on the total time.

The rat brain dataset introduced in Chapter 2 is re-used here (for a complete description and schematic see Figure 2.1).

4.4.1 Efficiency

All of the pipelines except the spatial correlation method were implemented in a memory efficient manner so that only a single spectrum was required in memory at a time. Most of the literature methods are have a stated aim or requirement of making the data small enough to fit into memory [66, 148]. For unsupervised

trend extraction using PCA multiple copies of the data are required in memory simultaneously (up to 4, depending on the algorithm used[174]), meaning that simply making the data smaller than the available memory may not be sufficient to enable further processing. Another requirement for further processing is that the measurements from each pixel are consistent, in a usual mass spectral storage scheme there may be differences in the exact m/z bins measured and as the data can be sparse often a reduced set of m/z-intensity pairs are recorded. All of the pipelines produced an internally consistent set of measurements where each pixel has values for each measurement.

4.4.2 Timings

Time is important in the processing of MSI and in general a processing time of less than the image acquisition time is desirable[2]. Disk access is the most significant time factor for all of the feature selection methods tested so the most pertinent measure is the number of data passes that are required, all the algorithms were implemented to be as pass efficient as possible, at any point that the data could be stored and accessed from memory it was. The data storage used (imzML, processed pairs) only permits easy access to individual spectra, some data storage frameworks allow for rapid access to whole ion images[203] but these are not commonly supported. As the code is implemented in MATLAB and is not optimised it is not really appropriate to make substantial commentary on the time taken but some relative times are provided for guidance.

It is clear from Table 4.2 that the disk load time is not the only consideration for timing, comparing the time for no processing with re-binning it is perhaps surprising that doing extra processing can reduce the total dimensionality reduction time, but using this non-optimised code the multiplication required for the BASC sampling step creates a large matrix in memory for every spectrum (of size $m \times k$), rebinning reduces the number of spectral channels from 129796 to 9500, which is a 93% reduction in sampling matrix size. During the evaluation it was discovered that it was possible to store the re-binned data in memory. A BASC implementation can be applied to data in memory which substantially reduced the computation time (to seconds, once the data is in memory) as a second data read could be avoided.

Memory efficient coding is not necessarily the fastest. For example, in the re-binning pipeline each spectrum is loaded and re-binned independently then the sampling matrix is formed spectrum-by-spectrum. Performing the re-binning as a separate stage and storing the result in memory allows faster matrix operations to be used for the basis construction and compression. This produced a final time of 0.06 (relative to Pipeline:none), including building the datacube. There is clearly a trade off between increased disk access and time but the disadvantage of algorithms that require data to be accessible in memory is that they will

not be useful for data that is very high-resolution in mass or space. A further consideration for future investigation is the ability to parallelise or distribute the computations in order to take advantage of cluster or distributed computing[194], the BASC approach lends itself well to parallelisation .

4.4.3 Dimensionality following Feature Selection

Part of the purpose of these pipelines is to reduce the dimensionality of the data and all succeeded in reducing the dimensionality of the data substantially, see Table 4.2, and thus making it more amenable to computation processing. The rebinning achieves a pre-determined level based on the bin width so this should be chosen based on knowledge of peak width or the final size of the data required. Within the peak-detection based methods there is a substantial difference in the number of peaks remaining between the most (summary spectra pipeline 5571 m/z s) and the least (intra-spectra 168 m/z s) as they have almost opposite inclusion criteria, the summary spectra aim to keep all peaks and amplifies small peaks whilst the intra-spectra keeps a few peaks that are seen in many pixels.

Name	Number	(% Original)
None	129796	(100)
Re-Binning 0.2	4750	(3.7)
Standard Deviation Spectrum	3440	(2.7)
Summary Spectra	5571	(4.3)
Intra-Spectra Top-Peaks	168	(0.13)
Spatial Correlation	669	(0.52)

Table 4.2: Number of m/z measurements retained following feature selection

4.5 Comparing the Data after the Pipelines

One indicator of discriminatory behaviour is the sparsity along the measurement[87] as highly discriminatory peaks will be present only in specific regions. Sparsity along the measurement vector is also an interesting measure as it measures the density of information contained within a vector. Another surrogate measure of information quantity is the eigenvalues of the principal components as they describe exactly how much variance exists within the data, and it is variance that provides discrimination.

4.5.1 Sparsity Measures

Sparsity of the feature selected is an interesting metric as it gives an indication of how 'information dense' resulting collection of measurements are. To determine if the processing affects this in a spatially biased

fashion the metrics were calculated pixelwise. The sparsity measures considered here all show more positive values as being more sparse.

Definition of measures of sparsity

l_n norms In general the l_n norm is defined as:

$$l_n = \left(\sum_{j=1}^{N} \mathbf{x}_j^n \right)^{\frac{1}{n}} \tag{4.1}$$

This norm is commonly encountered within mass spectrometry with $n = 1$ as the TIC which is the sum over all m/z channels within a spectrum (as m/z values are always ≥ 0, it is equal to the l_1 norm). To comply with the constraint that the sparsity measure gets more positive with increasing sparsity $-l_1$ is used (and its maximum value is 0).

pq-mean The pq-mean is defined as

$$r = -\left(\frac{1}{N} \sum_{j=1}^{N} \mathbf{x}_j^p \right)^{\frac{1}{p}} \left(\frac{1}{N} \sum_{j=1}^{N} \mathbf{x}_j^q \right)^{-\frac{1}{q}} \tag{4.2}$$

which forms a measure of sparsity when $p \leq 1$,$q > 1$[97], $p = 0.5$; $q = 1.5$.

Sparsity Following Feature Selection

Two measures of sparsity were used to investigate the output of the pipelines. The first, $-l_1$, was chosen as it is familiar within the community being closely related to the TIC metric. However, the $-l_1$ can fail to provide a robust measure of sparsity[95] so the pq-mean is also used.

$-l_1$ Norm The negative l_1 metric can be seen spatially in Figure 4.1 and can be visually interpreted as light areas corresponding to low ion counts and dark areas corresponding to high counts.

In the 'unprocessed' data (Figure 4.1.A) there is, broadly, a decrease within the tissue region and a significant 'checkerboard' patterning that is probably a matrix deposition artefact. The re-binning method (Figure 4.1.B) had no effect on this sparsity metric as it only redistributes ion signal between m/z bin spacings.

Peaking picking on the standard deviation spectrum (Figure 4.1.C) also produced a $-l_1$ metric almost identical to the original data, indicating that an equal number of ions are accumulated by this peak selection

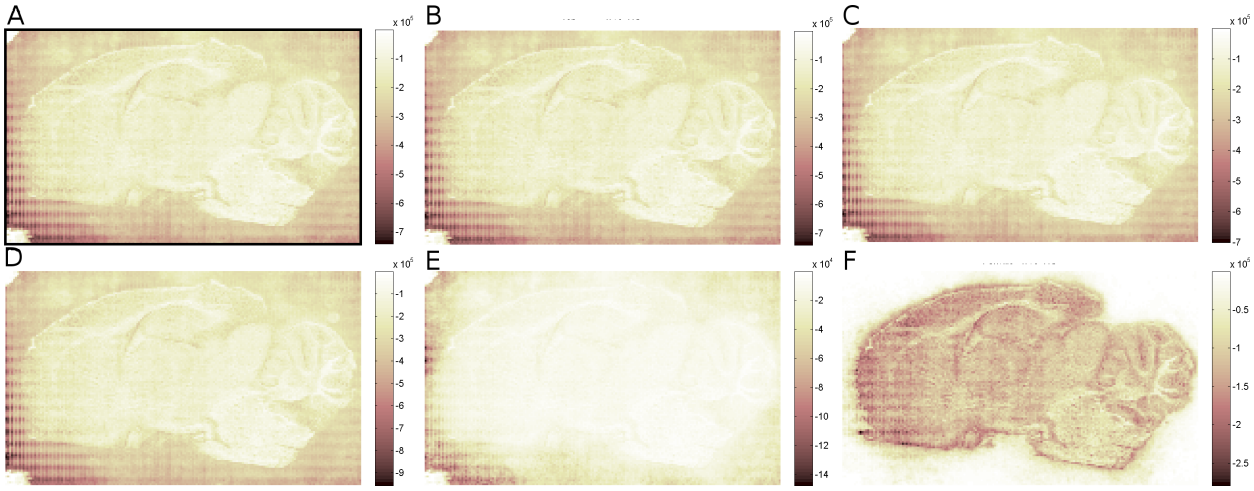

Figure 4.1: The $-l_1$ norm for each pixel shows the change in information density following processing. A. none; B. Re-bin 0.2; C. STD Spectrum; D. Summary Spectra; E. Frequent Peaks; F. Spatial Correlation).

method as were originally present despite reducing the number of measurement vectors to fewer than rebinning. Peak detection on a collection of summary spectra (Figure 4.1.D) produced a $-l_1$ with near identical *distribution* to the original but with substantially more negative values, indicating that individual ions were accumulated into multiple peak bins. This can happen as a simple filter window is applied during the datacube construction so, if the detected peaks are closer than the window, ion bins can be counted multiple times. Retaining inter-spectra peaks (Figure 4.1.E) returns a $-l_1$ metric that shows increased sparseness over the whole area but that is noticeably more sparse within the tissue region. Finally, spatial decorrelation (Figure 4.1.F) returns a $-l_1$ that is very sparse within the background region compared to the tissue area. The values for the on-tissue area are comparable to the same area in the original data suggesting that little is removed.

The value of using the $-l_1$ norm to assess sparsity is that it illustrates the total ion intensity. From this it is clear that only spatial decorrelation significantly shifts the relative data density towards the tissue area, which is what this pipeline aims to achieve.

pq-norm It is useful to use the pq-mean as a second sparsity metric as it is not vulnerable to simply shifting intensity bins, like the l_1 norm was seen to be. As a ratio of vector norms it is insensitive to intensity scaling but is more sensitive to 'spikiness' within the data. It returns zero when all elements are an equal value and tends to zero for a single data spike[95]. The pixelwise calculation of this metric is shown in Figure 4.2.

For the unprocessed data (4.2.A) the sparsity is uniform over the imaging area with a few patches, mostly

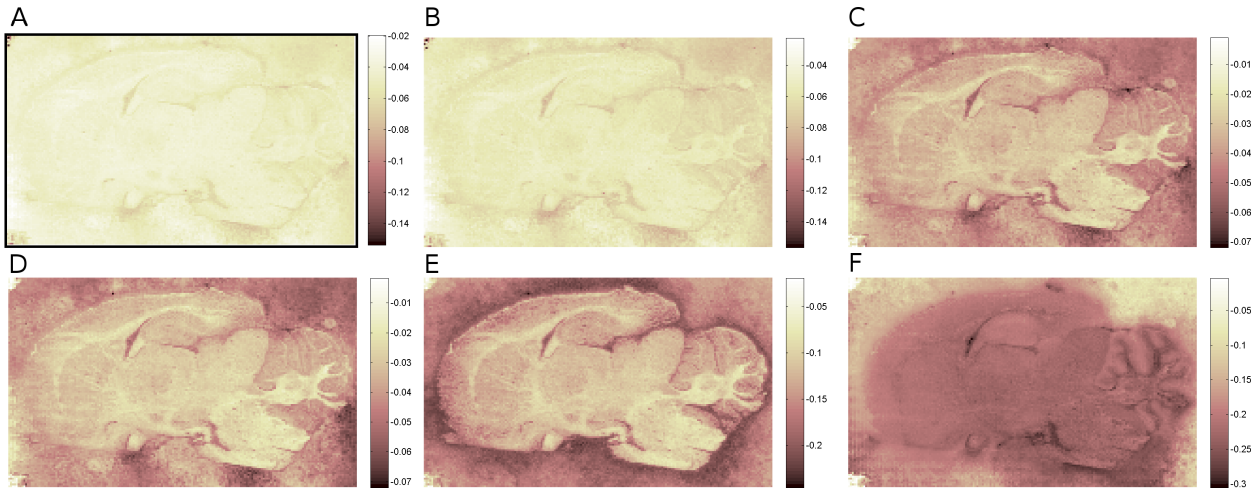

Figure 4.2: The pq norm shows the data sparsity following each pipeline. A. none; B. Re-bin 0.2; C. STD Spectrum; D. Summary Spectra; E. Frequent Peaks; F. Spatial Correlation).

at the tissue border, that are relatively dense, contrast this with the $-l_1$ norm that showed significantly lower sparsity in the off-tissue areas. The pq-mean metric indicates that peak-density of both areas is clearly similar whilst the $-l_1$ suggests that there is higher magnitude off tissue (Figure 4.1). Re-binning has some subtle effects on the pq-mean (Figure 4.2.B) with some regions, mostly off tissue, become less sparse, perhaps suggesting that peaks are becoming merged during the re-binning, but the effect is slight. There is a substantial change visible in the summary spectra peak picking examples (Figure 4.2.C-D) where overall sparsity substantially increases in every area. In both cases the off-tissue regions become less sparse compared to the on-tissue and anatomical variations become apparent. Anatomy is also very visible in the pq-mean following inter-spectrum peak selection (Figure 4.2.E) and the tissue boundary is particularly clear, however, overall the sparsity is substantially decreased suggesting that this method may be detecting common peaks with more similar values. Sparsity is further decreased following spatial decorrelation (Figure 4.2.F) with the lowest values seen on tissue. Little anatomical variation is present and the background is very sparse.

The combination of sparsity metrics provides an insight into the data, which would be difficult to see from manual interpretation. Quality metrics for mass spectrometry images are a subject of current development[3] so it will be interesting to see whether these metrics correlate with segmentation performance and whether they could contribute to an unbiased estimation of data quality.

4.5.2 BASC-PCA

To enable a fair comparison to be made between the different pipelines, factorisation using PCA was added as a final feature step to each of the pipelines. The main motivation behind this additional step was to provide a uniform input to the visualisation with a well understood statistical approach. Additionally, the PCA loadings provide further insight into the differences in data features retained between the pipelines and the variance per component is produced.

Variance

PCA concentrate the variance that exists within the data and the components it produces are ordered by the amount of variance they contain. Figure 4.3 shows the absolute variances for the first 10 components and what fraction this is of the total variance from each pipeline. The data-cube based pipelines generally

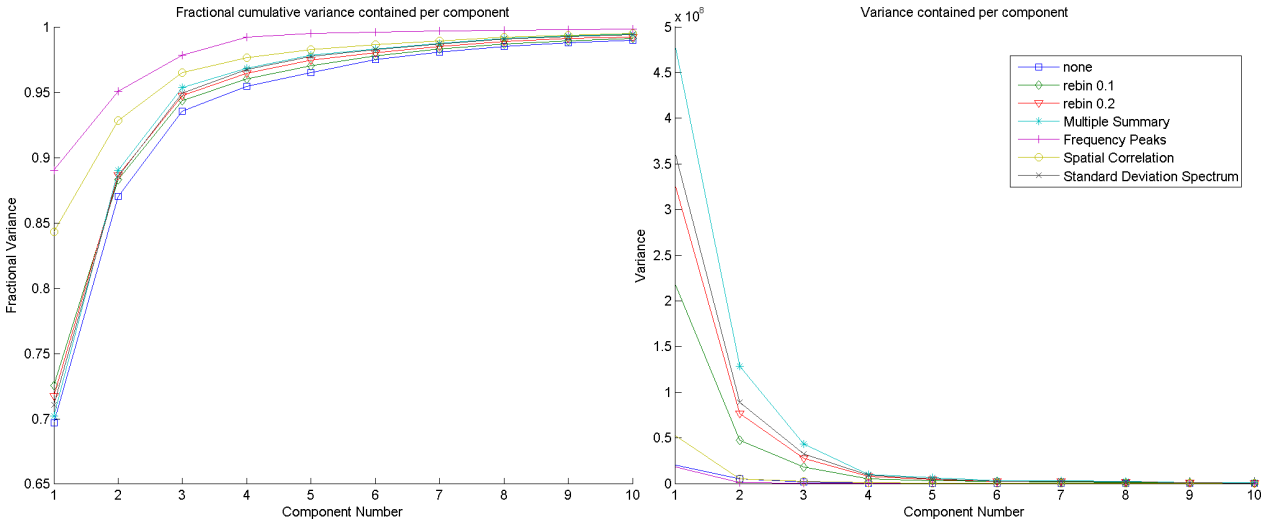

Figure 4.3: The variance contained in PCA component following each pipeline. right: absolute variance. left: cumulative fraction of variance.

had greater magnitudes of variance as each channel has larger magnitude (due to it being the sum of several channels) and so the re-binned datasets have greater variance than the original. The more selective pipelines (spatial correlation and frequent peaks) had the lowest total variance, presumably due to the removal of signals from unwanted channels. These pipelines also had a greater fraction of the variance contained in the first 5 components, suggesting that many of the measurements preserved by these pipelines had similar spatial distributions.

PCA Score Images

As described in Chapter 3, PCA concentrates measurements so that variables that follow the same trends can be visualised as a single map. The first three principal components loading maps are shown in Figure 4.4.A. As a broad summary, for the unprocessed data the first component separates tissue from off-tissue, the second component makes a 'halo' of leached lipids visible and the third component distinguishes between grey and white matter. This trend holds for the re-binned and summary spectra methods Figure 4.4.B-D. This is unsurprising for the re-binned and suggests that both peak picking methods make an effective and unbiased summary of the data. That the multiple-summary method picked some peaks more than once (as was noticeable from the $-l_1$ norm) does not seem to effect these results substantially, as these measurements are presumably combined by the factorisation. The intra-spectrum peak picking Figure 4.4.E produces a similar first component to the previously discussed methods but components 2 & 3 appear to be in the reverse order so more tissue detail is visible in component two.. Following spatial decorrelation the tissue features are much more pronounced in all of the first three components (Figure 4.4.F). As can be seen from the colour-bars, the colour-scheme has to be adjusted in some cases for clarity of visualisation, particularly in the case of the decorrelated images.

4.5.3 Segmentation

The use of a SOM is a suitable visualisation of the feature extraction results as it provides a single map describing the data with coloured graduations. As the SOM automatically scales to encompass the variations retained following feature selection it can deal with differences in data range due to different numbers of measurements preserved. The orientation of the SOM within the data space is randomly determined during initialisation so the colourmaps were rotated (in node-space) by a multiple of 90 degrees so that the same spatial regions had similar colours, but no scaling or other adjustments to the colour scheme was made. Figure 4.5 shows the SOM segmentation of the whole image scene. Using a 2D SOM, with four colours, inevitably emphasises the four extremal parts, in the case of the unprocessed data (Figure 4.5.A) these highlight two regions within the matrix-background and separate the grey and white matter on-tissue. Anatomical features are apparent within the on-tissue area and the 'checkerboard' artefact is also visible. As the SOM is limited in colour dynamic range it was possible that more spectral differences within tissue anatomy could be visualised if the spatial area was limited to only the on-tissue region during the application of the SOM (i.e. the full dataset was used for the feature selection and extraction pipelines and the BASC-PCA). The SOM output for the masked data is shown in Figure 4.6. There is an immediate improvement in the anatomical detail visible

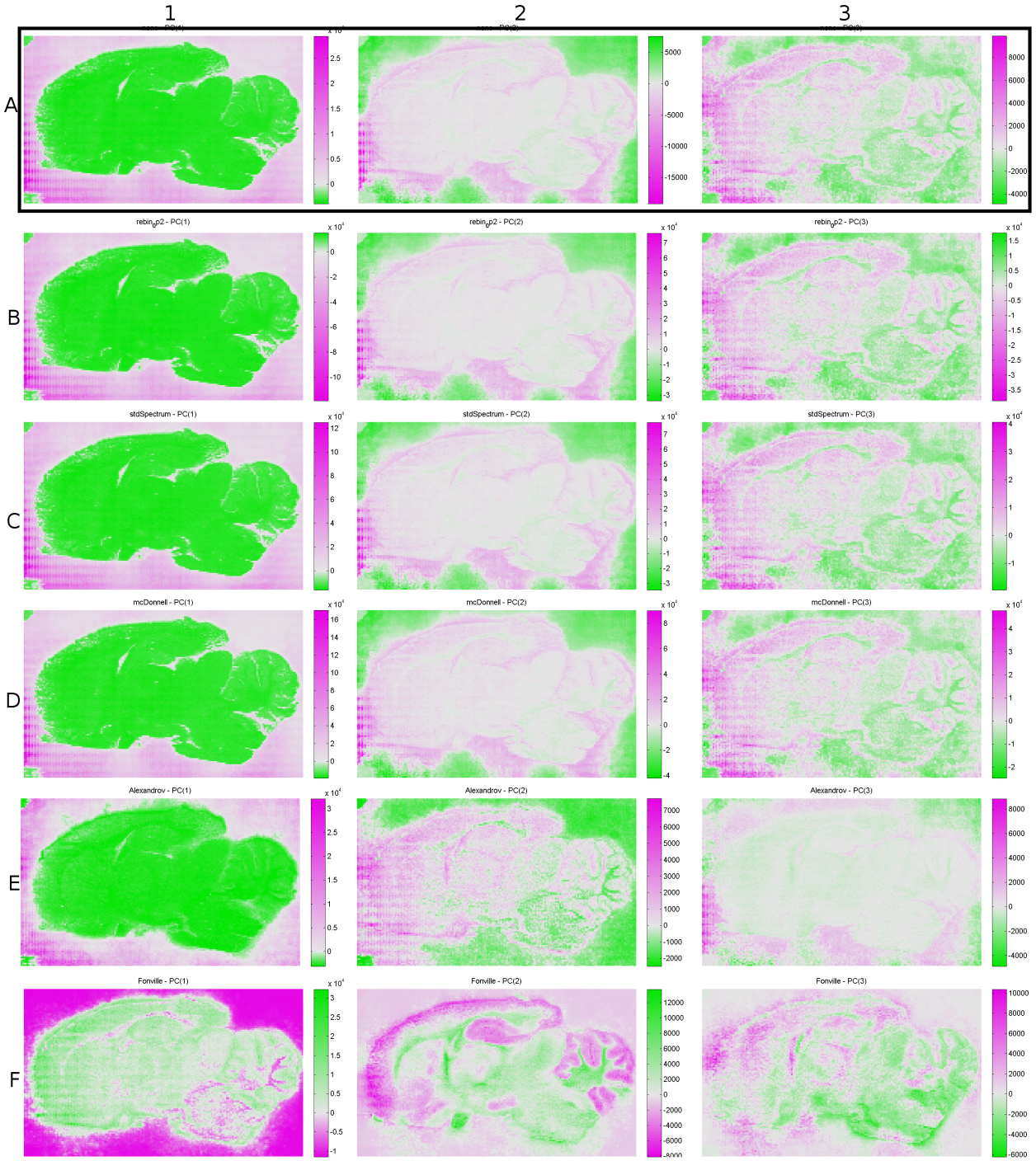

Figure 4.4: Principal components 1-3 following the feature selection pipelines. A. none; B. Re-bin 0.2; C. STD Spectrum; D. Summary Spectra; E. Frequent Peaks; F. Spatial Correlation).

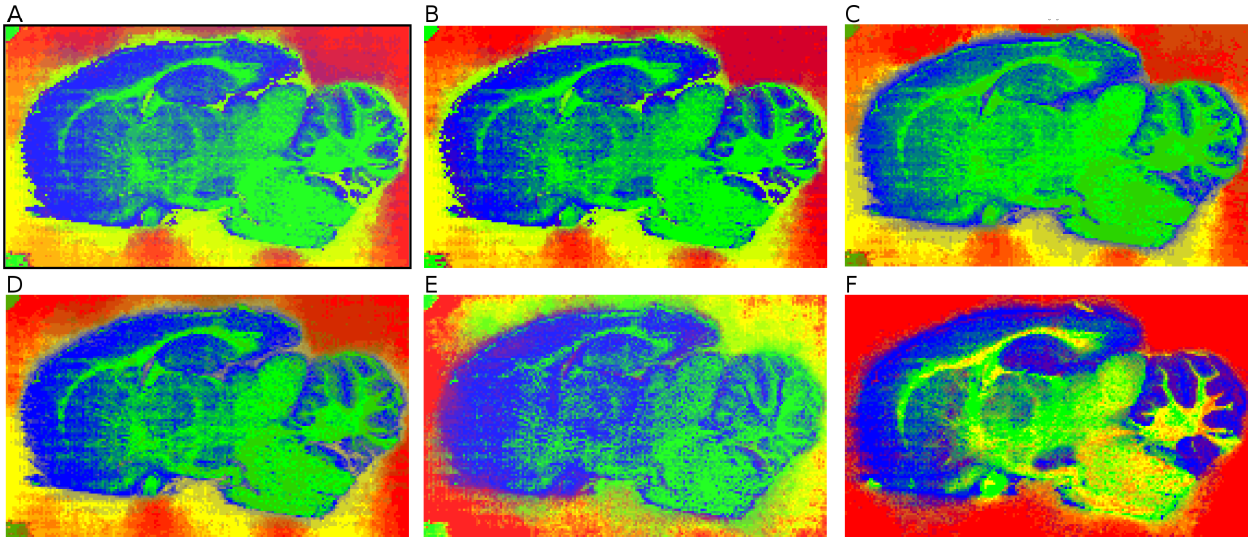

Figure 4.5: Self organising maps used to segment the fixed rat brain image following each pipeline. A. none; B. Re-bin 0.2; C. STD Spectrum; D. Summary Spectra; E. Frequent Peaks; F. Spatial Correlation).

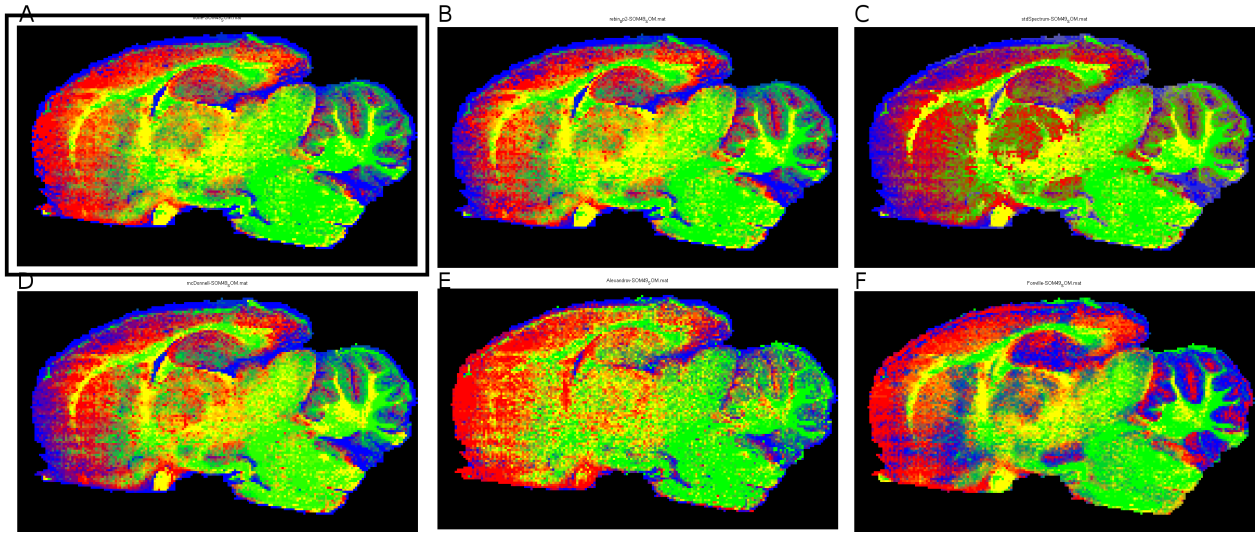

Figure 4.6: Self organising maps with a tissue mask. The mask was applied after feature selection and feature extraction using PCA to provide the SOM with a greater colour range. Compared to Figure 4.5 this allows much more details to appear in the tissue region. A. none; B. Re-bin 0.2; C. STD Spectrum; D. Summary Spectra; E. Frequent Peaks; F. Spatial Correlation).

in all of the segmentation maps, highlighting how important it is to match segmentation dynamic range to the features of the data. This yields substantially less difference between the features visible following spatial decorrelation (Figure 4.6.F) and the other methods. The 'none' and rebinned segmentation maps look almost identical indicating that little in the way of noise suppression is achieved by re-binning but also that little discriminative information is lost. Some additional structure (yellow) is visible in the cerebellum. Using the standard deviation spectrum or a cohort of summary spectra retains all the anatomical features visible in the raw data. The approach of peak filtering separates some of the major anatomical structure but fails to retain all of the details, particularly within the cerebellum. Spatial correlation returns a very similar anatomical pattern with substantial differences in the classification of the hippocampus and cerebellum. This illustrates that the choice of pipeline can have an impact on the final digital histology.

4.6 Conclusion

The goal of examining different feature selection and extraction pipelines was accomplished by comparing several evaluation metrics against a baseline of BASC on raw data. It is only using the memory-efficient implementation of basis approximation that it was possible to generate a baseline data-set and so explore the relative effects of the pipelines. Specific attention was paid to memory efficient algorithms for this comparison as they have a distinct advantage in that they can be applied to data containing any number of spectra. There is even the potential that they can be run in parallel with data collection so a substantial portion of processing can be done whilst the data is still being collected.

The comparison of several pipelines revealed that, by-and-large, peak picking on summary spectra preserved the spectral and spatial features of the original dataset. Using a method that actively penalised m/z distributions that correlated with off-tissue distributions did result in increased tissue contrast in subsequent factorisation and segmentation. However, simply masking these areas prior to segmentation improved the pipelines that did not incorporate spatial processing steps. This suggests that if prior knowledge of the background area is available then excluding it from the measurement region reduces the complexity of subsequent analysis. If rebinning is performed then this should be carefully tuned to mass resolution of the instrument but if performed on-line with data compression can be implemented so that an additional data-load can be avoided. The standard deviation spectrum for feature selection follows the template of other summary spectrum picking schemes[148, 157] but naturally eliminates any channels that have zero variance and are thus inherently uninformative.

These results were generated from a single dataset collected from an instrument with a intermediate

mass resolving power. This approach of using basis approximation as a baseline could be applied to other datasets, particularly very high mass resolution images, to provide a baseline for understanding effects of data processing. It would be recommended that any new data processing pipeline be compared to the results from unprocessed data as using basis approximation for spectral compression there is no longer a restriction on the size of raw data that can be evaluated. One interesting test case for this could be the statistical artefact removal developed by Gerber et al[73] to remove 'tiling effects' that occur in Secondary Ion Mass Spectrometry (SIMS) (which are visually similar to the checkerboard patters seen in the test dataset used in this chapter). This could be used as a check to ensure that no other image features are removed during processing.

A significant challenge highlighted during this work is the difficulty in establishing measures of the differences between algorithms, they can be compared for parity but when differences occur determining what is 'best' can only be qualitatively described. There is a need for the generation of baseline datasets with well described underlying spectral characteristics for the evaluation of the multiple methods.

Chapter 5

Evaluating Data Processing using Simulated Mass Spectrometry Imaging (MSI)

In earlier chapters the improved efficacy of existing segmentation approaches to mass spectrometry images was demonstrated using basis approximation to compress the spectral data. This provides a route to establishing a quantitative estimation of the sensitivity and specificity of data processing pipelines using a known ground truth. For the evaluation of data visualisation algorithms the composition of the ion packet at each point on the sample surface must be known. The simulation takes a set of ion lists with defined spatial distributions and a statistical model derived from a mass spectrometer and returns a dataset that closely resembles realistic spectra.

The simulated data was used to quantitatively evaluate the success of segmentation with k-means and spectral clustering, and optimise the clustering parameters. For the segmentation of MSI data spectral clustering was found to be a more robust method than the classic k-means algorithm. Additionally, this provides a method for distributing standard datasets for evaluation by the community.

5.1 Introduction

Automated approaches for the visualisation and interpretation of the large spatially and spectrally heterogeneous samples produced by MSI have been discussed at length within this thesis. This has lead to the development of several new techniques for this purpose within Chapters 2 & 3 which expand the methods already available in the literature (e.g. [4, 52, 67, 108, 146]). Qualitative methods for the assessment of segmentation by automated pipelines were discussed in Section 1.6 where it was concluded that the typical evaluation criteria of visual comparison is a limited and naive approach that is unable to provide a quantitative metric for comparing algorithms.

To obtain an absolute metric algorithms must be applied to a dataset where every spectrum has a well characterised spectral composition, a known 'ground truth', to which the algorithm output can be compared. There are no such comprehensively characterised datasets currently publicly available for MSI, probably as a result of the characterisation of tissue samples being notoriously difficult[89] and the data size making transmitting and sharing data a challenge[2]. An ideal imaging test set for the evaluation of image processing pipelines would include annotations at every spatial location with a description of each spectral feature present. With these details included a fair comparison of the information extraction abilities of algorithms could be made. For mass spectrometry imaging these annotations could include ion composition and defined distributions against a background of statistically representative experimental and biological noise.

One method of producing such a well characterised dataset is to simulate data that maintains all of the characteristics of experimental images but is produced from pre-determined molecular and spatial patterns. The use of simulation for evaluating spectral feature selection algorithms has been established for individual Time of Flight (TOF) spectra[43, 157] and for LC-MS and LC-MS/MS spectra[176] on a physics-based model of the individual instrument. In an imaging context, mixtures of average spectra with added Poisson noise have been used to form test images[87, 214, 217]. Poisson noise is an important characteristic of mass spectrometry data but chemical noise and matrix variation provide substantial disruptions within real signals[193]. A more holistic approach is presented here which includes accurate peak shapes, varied spatial patterns with instrument and chemical noise applied. These descriptions may be randomly generated or based on known molecules ionised, the molecular mass, adduct formed and absolute abundances, alongside a model for the spectrometer that allows the incorporation of both chemical and instrument noise that is able to produce datasets of a realistic size.

5.1.1 Instrument Response Function

This work takes an Instrument Response Function (IRF) modelling approach to generating spectra. A series of IRFs are generated that describe the effects that the components within the instrument have on the signal, providing a statistical modulation of a pure input spectrum. Using IRFs, which can be experimentally determined, provides a numerically straightforward method for simulating realistic ion data. The IRFs can be calculated analytically from physical laws, experimentally from measurements of real data or through simulation using a sophisticated physics based model, these can be used separately or in combination for a full approximation of instrument characteristics [43, 170, 177].

To allow for maximum flexibility a set of IRF signal modifying terms are considered. These fall into two categories: those that affect the mass to charge ratio (m/z) bins independently (additive and scaling factors) and those that affect a range of m/z values. The additive term is generally noise related (e.g. electronic or shot noise), the scaling term relates to ion transfer efficiencies, and the shaped filter models the spreading of ion signals into the characteristic mass spectrometry peaks.

The route taken to establishing the IRF terms to include, and their parameters, is central to the results presented here. In general the individual components are assessed for their likely IRF properties based on literature or intuition, the specific parameters required are then determined from experimental data. The model parameters will be established on the example of a QqToF instrument but the process will be explained with a view to generalisation as much as possible. The aim of this simulation is to separate the effects of ion generation from instrument noise in order to evaluate the ability of algorithms to extract patterns from the ions. As the current understanding of Matrix Assisted Laser Desorption Ionisation (MALDI) ion generation is not sufficient to predict which ions will be seen from a particular complex mixture the simulation presented here will start *after* ionisation (with a peak list). Using the simulated data, a demonstration of the optimisation and evaluation of spectral clustering will be performed, so that pattern extraction from unknown data can be performed with confidence.

Additionally, this mass spectrometry image simulator could provide researchers with a tool that can reliably produce data of a known character based on known spectral profiles with defined spatial extent. The key to understanding how much the data processing impacts research conclusions is to have shared data and well described methods. Such a simulated dataset could now be used as a testbed for inter-laboratory data processing comparisons separating this issue from the added variation produced by sample preparation protocols.

5.2 Overview of Instrument Modelling

The approach to instrument modelling used here differs from previous work which started with a physical instrument model and built an analytical description of the peak output[43, 100], instead IRFs are generated from a specific instrument by fitting mathematical functions that directly describe the data. This allows the process to be applied to an arbitrary instrument for the comparison of the same benchmark dataset across platforms. The processes involved in approximating an individual spectrum will be explained followed by the extension of these concepts to an entire mass spectrometry image and this is illustrated for a single spectrum in Figure 5.1.

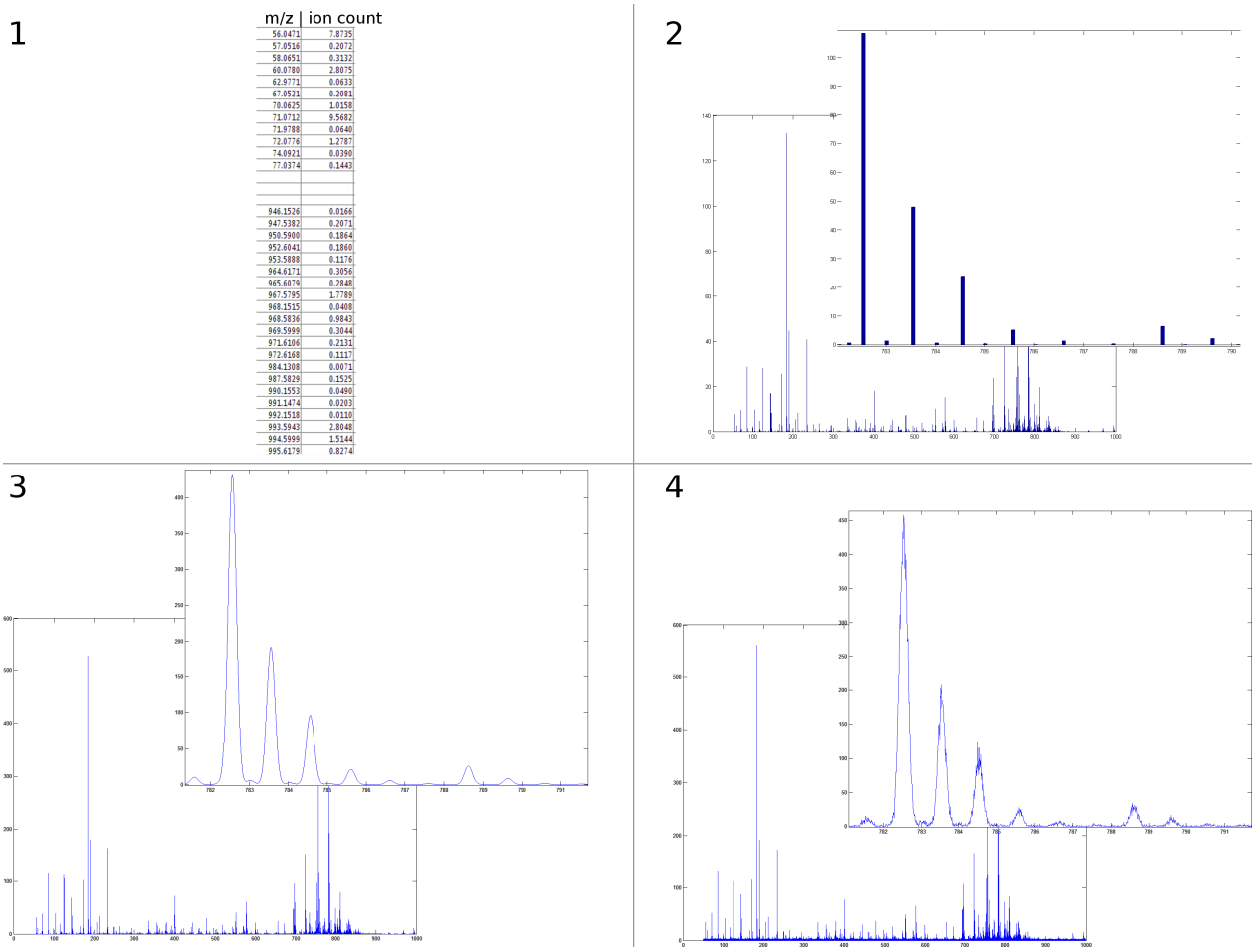

Figure 5.1: The key stages of simulating a spectrum (single pixel) 1. Input list of peaks and counts 2. Interpolated onto m/z axis defined by m/z range and detector resolution. 3. Convolution provides a blurring into a peak shape 4. Noise added (poisson and Gaussian) distort peaks

5.2.1 Spatial Distributions

The simulation is capable of generating MSI datasets that feature complex, overlapping spatial features. These are input as multiple layers, illustrated in Figure 5.2, each of which consists of a layer specific ion list and an intensity distribution. The distribution map values are a [0 1] scaling factor for each pixel. The final ion list for a pixel is the aggregate of the ion list in each layer multiplied by the scaling factor. A single ion list per pixel is then passed to the next stage of the simulation.

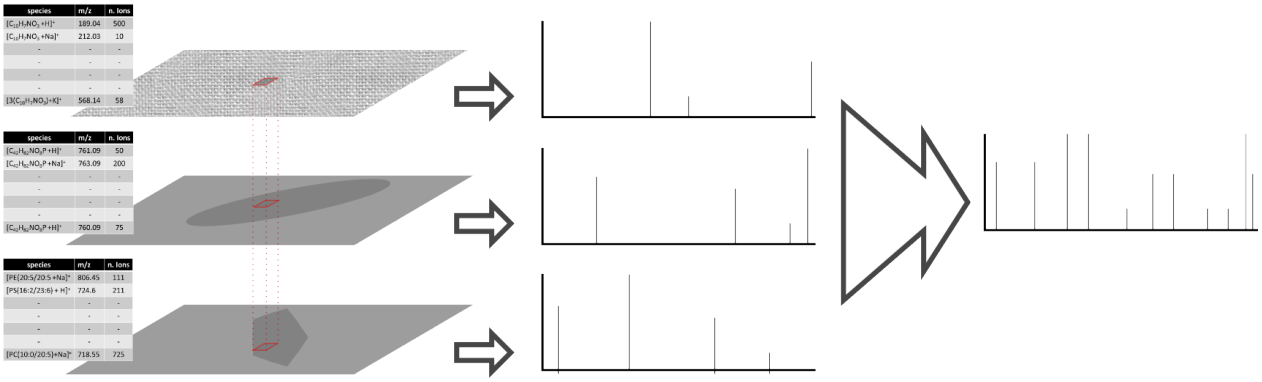

Figure 5.2: Each layer within th simulation has an ion list with a relative count. For a given pixel, this list is modulated by an abundance map. Pixelwise, the modulated list is interpolated onto the instrument m/z domain to give a spectrum per layer, these are added to give a 'pure' spectrum for each pixel.

5.2.2 Spectral Input

The starting point for the simulation is an initial list of ions present at each pixel location. Each ion list (Figure 5.1a) consists of pairs of exact mass values and ion intensities. This list could be randomly generated; an extracted set of m/z values; or calculated from the sum formula of known molecules (with isotope patterns and optional adducts). Note that this simulation starts *after* ionisation which allows the resulting instrument model to be appended onto any ion source model, including other ionisation types such as SIMS or DESI.

5.2.3 Discretising the m/z domain

The physical process of mass analysis generates continuous values along the m/z axis which are then divided into discrete mass bins by the detector so all ions that are detected within a $\Delta m/z$ of a particular bin are aggregated. The final mass spacing is defined by both the type of instrument analyser and detector[93] alongside the experimental configuration. During modelling the edges of the mass bins are determined and the input ions are allocated to the bin that contains their m/z value, summing the intensities of ions that

become isobaric. This populates a mass axis binned with the appropriate mass resolution that is then passed to IRF functions (Figure 5.1b) which operate on the assumption of regularly spaced mass intervals.

5.2.4 Choosing IRF Terms

The types of IRF are described in general here and the process of choosing the correct parameters for IRF functions for a specific instrument will be demonstrated in Section 5.3. Care must be taken with the order of IRF operations as subsequent terms are applied to previous results, e.g. signal blurring due to the mass analysers must be modelled before detector noise terms are added. The signal at a particular m/z bin can be modelled by two types of function: those which are influenced by intensities of neighbouring m/z bins and those that affect each m/z bin independently.

Neighbour-Effected Functions

Kernel based filters move a shaped function along the m/z axis evaluating it at each bin[188], so the value at a particular m/z is dependent on the values in the immediate spectral neighbourhood. A suitably chosen filter superimposes an ion peak shape onto the pseudo-delta function of the input signal to simulate the 'blurring' that arises due to the measurement process. Figure 5.1c illustrates the effect a Gaussian filter has on the pure signal: transforming it from spikes to wider peaks.

Filter Implementation with Convolution Convolution is regularly used in IRF modelling as it is a rapid method for applying a filter to the whole signal using a Fourier transform [29]. As it applies the same filter to every measurement simultaneously it requires a linear measurement spacing so it is important that the m/z domain is transformed if required.

m/z Independent Functions

Additive terms are applied to each bin independently on the filtered signal but can be a function of the m/z value. Some examples of independent effects reported in the literature include the baseline artefact, seen within linear ToF instruments caused by saturation of the detector by low molecular weight fragments which is commonly modelled as an exponentially decaying additive term[226]. Electronic noise which arises due to the thermal motion of electrons within the detection circuitry can result in an additive component drawn from a Gaussian distribution[43]. Poisson or shot noise is experienced by all counting-type detectors, the magnitude of which is proportional the square root of the intensity of the signal[60]. Chemical noise considers

the addition of non-spatially correlated peaks to represent genuine detections of molecules that are randomly distributed. It is not strictly an IRF but may be useful for the evaluation of the robustness of algorithms, this must be added before any convolution in order to have a suitable peak shape[122].

5.3 Modelling the QStar Elite Instrument

The parameters for a QStar Elite QqToF (AB, Warrington, UK) instrument were determined by analysing the fixed rat brain MALDI MSI dataset introduced in Chapter 2. The QStar Elite was operated with the quadrupoles functioning as an ion guide with collisional cooling and an orthogonal reflectron time-of-flight (QqToF) path directing the ion onto a Micro Channel Plate (MCP) detector with an time to digital converter (TDC) providing a m/z range of 50-1000. Following the general stages outlined in Section 5.2 and Figure 5.1, each component of the instrument (post-ionisation) was considered and IRF model functions designed to mimic them were created:

5.3.1 Instrument Modelling from a MSI dataset

MSI datasets are useful for characterising an instrument as they provide many measurements collected over a short time period, so external sources of variation are minimised. Due to the large number of spectra, and the many hundreds of peaks within each spectrum statistical models can be fitted to the measurements that describe both the intra-spectrum features and the inter-spectra noise. The downside of using a MSI dataset is that there is an unknown heterogeneous sample which may introduce errors in the spatial modelling. Having a complete statistical model for a particular instrument allows datasets with realistic characteristics to be generated and used to evaluate data processing methods. These can then be applied with confidence to further data collected from that instrument.

5.3.2 Discretising the m/z axis

Time-of-flight mass analysis records a signal in the time domain are then transformed to m/z values. This introduced a non-linearity in the spacing along the m/z axis which is a function of both mass analyser and detector. The final m/z axis can be calculated as follows:

Mass Analysers

Two mass analysers are present in the QStar-XL, a set of three quadrupoles and the TOF. During the imaging experiment the quadrupole voltages were optimised to function simply as an ion guide and so are assumed to have negligible effect on the ions in the mass range considered[93], consequently they are not modelled. Having the TOF pulser oriented orthogonally to the ion source (combined with collisional cooling in the quadrupoles) physically decouples the mass analysis from ion source effects like sample topology and surface charging[84] so the physical process of TOF mass analysis can be considered independently.

Ions were accelerated with a pulsed voltage, V, into a field-free region (the drift tube), as each ion has the same kinetic energy, $E_k = zeV$ (where m is the ion mass, z is the charge on the ion and e is the charge of a single electron), they achieve a different velocity, v according to[93]:

$$E_k = \frac{mv^2}{2} = zeV \tag{5.1}$$

$$v = \sqrt{\frac{2zeV}{m}} \tag{5.2}$$

So, the time, t, taken for a singly charged ion to travel a distance L is

$$t = \frac{L}{v} \tag{5.3}$$

$$t^2 = \frac{m}{z}\left(\frac{L^2}{2eV}\right) \tag{5.4}$$

$$\frac{m}{z} = ct^2 \tag{5.5}$$

Mass Detector

The detector time resolution determines the final discretisation data and records it with a fixed time frequency, δ_t, so the bins are equally spaced in time. The non-linearity in the m/z domain, and linearity in the time domain, is shown in Figure 5.3. Knowing this and Equation 5.5 allows us to write the final m/z axis as the following equation[93]:

$$\mathbf{m} = [ct_0^2, c(t_0+\delta_t)^2, c(t_0+2\delta_t)^2, c(t_0+3\delta_t)^2 \ldots c(t_0+n\delta_t)^2] \tag{5.6}$$

where n is the total number of m/z bins.

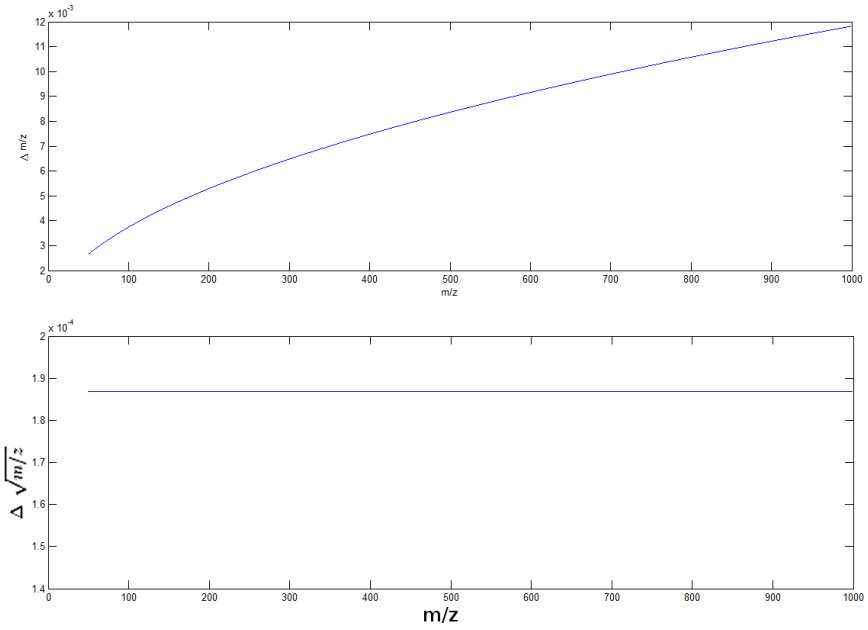

Figure 5.3: Calculating the difference between m/z bin centroids. Top. the difference is non-linear in the m/z domain Bottom. The $\sqrt{\mathrm{m/z}}$ bin difference is a constant: (δ_t).

Generating a m/z axis

The bin spacing (including the calibration coefficient),$c\delta_t$, was read directly from Figure 5.3 (from the graph in the time domain: $\sqrt{\mathrm{m/z}}$) as $c\delta_t = 1.875\mathrm{e}^{-4}$. Several spectra from the dataset were examined to confirm that this axis was constant over the whole experiment. The starting value for a spectral axis $ct_0 = \sqrt{\mathbf{m}_1}$ The m/z axis can be generated on-the-fly based on a minimum and maximum value, m_1 and m_{end} using Equation 5.6, which is useful for adjusting the m/z range. This also means that the mass inaccuracy due to discretisation increases as a function of the square root of the m/z.

5.3.3 Peak Shape

Ions are pulsed into the TOF analyser at right angles to their path out of the quadrupoles (see Figure 1.2). Collisional cooling within the quadrupoles has transformed the pulsed packet of ions produced on the sample surface into a narrow ion beam within which the ions have similar momentum[186]. The high frequency pulsing (around 10kHz for small molecule ions[186]) samples from this ion beam preserving the natural distribution of position and momentum orthogonal to the beam path. As the ion beam current is small and the pulsing frequency high the number of ions per pulse is low. The ions drift through the flight tube as described in Section 5.3.2 until they impact the MCP. An incoming ion from a pulse excites an

electron shower within the charged plates of the MCP which then fall on the TDC causing a voltage spike. Providing the voltage spike is higher than a threshold (100mV on the QStarElite [186]) a discrete 'count' is recorded and stored in a hardware bin. The detector then has some small delay during which it is unable to record another pulse (around 5 ns). Many pulses are performed and the final spectrum read out. Providing that the pulsed ion current results in less than one ion per dead time the normal distribution is preserved through detection and the final peaks will have a Gaussian profile (more advanced models can account for some detector saturation[214]):

$$f(t) = \frac{1}{\kappa\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \tag{5.7}$$

where the variance term, κ, determines the peak width and the distribution mean, μ, is the ion's mass (the peak is symmetric around the peak centroid m/z). The function is parametrised by t as it is applied to the time domain (the square root of the m/z).

It is known from literature and simulation that the peak width (controlled by the variance σ in Equation 5.7) increases and maximum peak intensity decreases as a function of m/z but that peak area remains constant[43]. The normalised Gaussian has unit area so it takes into account the relationship between peak width and maximum peak intensity but a single convolution does not allow for the peak width to vary.

In the case where multiple analysers are present (e.g. an ion guide preceding a time-of-flight analyser) their effects can be merged into a single experimentally determined convolution.

Fitting a Gaussian to spectral peaks

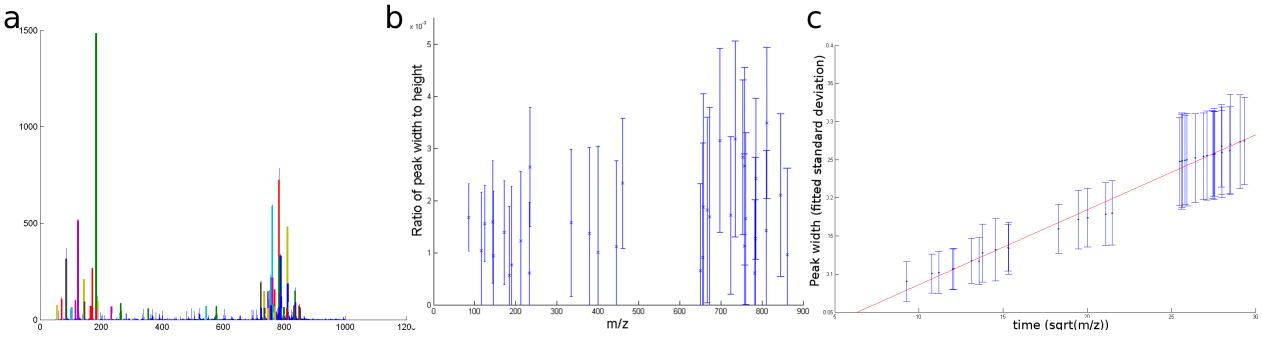

Figure 5.4: Profiling the statistical properties of peak shapes across a MALDI MSI dataset a. 35 peaks selected from the mean spectrum for profile evaluation which cover the mass range considered. b. plotting the ratio of peak height to width shows there is no m/z dependent relationship. c. Average peak standard deviation for all peaks shown in the time domain ($\sqrt{m/z}$) shows a linear relationship between standard deviation and time.

To model the changes in peak width over the dataset and m/z range a list of the most intense 35 peaks

was made from a mean spectrum, see Figure 5.4, that cover the mass range considered. For an instrument of this resolution (mass resolving power 6000 @ m/z 700) it cannot be guaranteed that isobaric ions are not present within a single peak. However, we assume this to be the case for most ions and trust that averages over a number of intense peaks remove any error due to the merging of peaks. From these measurements a good estimate for function parameters can be made.

The image was traversed pixel-by-pixel and for each of these peaks a Gaussian distribution was fitted by iteratively minimising the Root Mean Square (RMS) error between the Gaussian estimate and the data. The peak width is largely independent of the apex height, see Figure 5.4, but as expected[43] there is a broadening of the peaks as m/z increases which is linear in the time domain (i.e. $\sqrt{m/z}$). The fit generated parameters of standard deviation, peak height and peak area. The standard deviation controls the peak width, the full width half maximum (FWHM) is FWHM $\approxeq 2\sqrt{2\ln 2}\kappa \approxeq 2.354\kappa$ The relationship between κ and $\sqrt{(}m/z)$ was obtained by fitting a straight line to Figure 5.4c giving $\kappa = 0.01m/z - 0.001$.

Implementation with Convolution

Convolution applies a shaped filter, f (such as the Gaussian peak shape) to a vector $\mathbf{x}$. $\mathbf{x}' = \mathbf{x} * f$, effectively moving the filter along the vector and calculating the joint area at every point. As the filter is a function of the bin index $f = f(i)$ it is slow to apply this directly to a spectrum as the filter must be recalculated at every point. If a constant filter is used then the convolution can be performed very quickly via the fourier transform[23]. As a compromise to enable MSI datasets with tens of thousands of spectra to be produced the m/z axis is split into portions and a fixed width filter applied to each. As the filter width changes quite slowly over the range considered 5 portions were heuristically determined to be sufficient.

Coombes et al[43] asserted that the reciprocal of the peak height varied linearly with m/z. There was found to be a quadratic fit using the integrated convolution approach, see Figure 5.5, but detailed comparison to Figure 6 from[43] also showed similar periodicity in the residuals from their linear fit. In both our case and the published work the trend is dominated by the first order linear relationship even if in general a quadratic fit may be more appropriate.

5.3.4 Noise

The majority of components within a mass spectrometer operate independently and so their IRFs can be applied sequentially. This is true for the components of the linear TOF and QqTOF systems described as examples here.

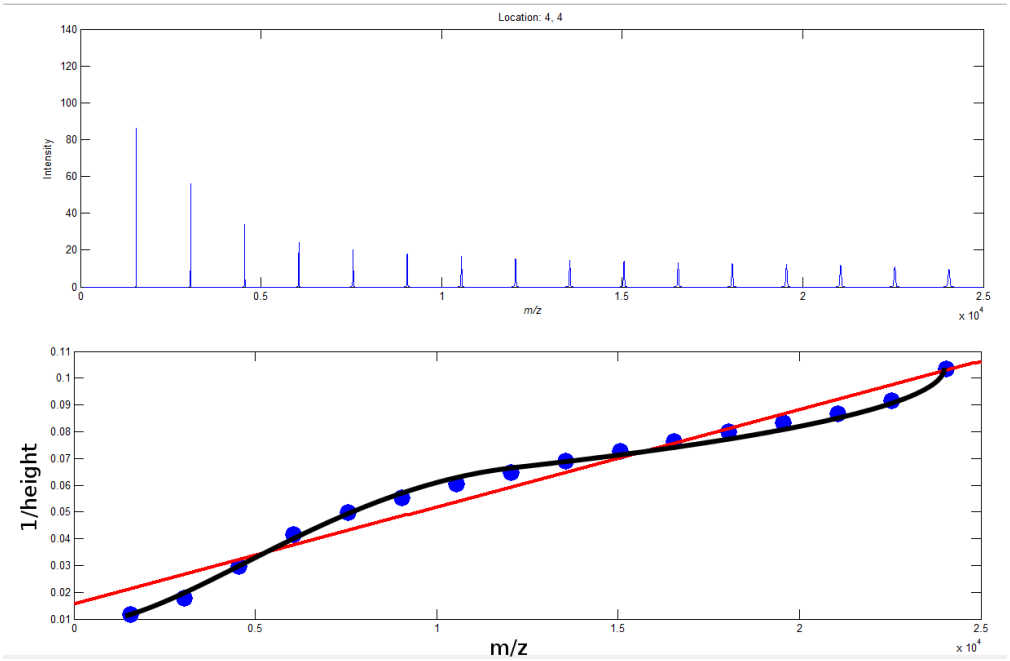

Figure 5.5: A simulated spectrum across a wide mass range. Top. Each peak was defined with the same number of ions at m/z [50:1500:25000]. Bottom. The reciprocal of the height was described as an approximately linear function in previous simulations [43] (red line) but here a slightly quadratic response was seen (black line). Subsequent re-inspection of [43] also showed some deviation from a linear fit.

Independent

The noise originating from thermal motion of electrons in the detector electronics is modelled with a small amount of additive Gaussian noise. Thermal noise is well characterised by a Gaussian profile[43, 214]. However, the TDC type detector effectively implements a hardware threshold on intensity values of less than 100mV (≈1 ion count [186]) which minimises the effect of electronic noise but also prevents a direct measurement. A heuristic noise standard deviation of $\beta = 5\%$ of a single ion count was therefore chosen. For instruments where substantial thermal noise exists it could be estimated from a region with no peaks.

Intensity dependent

As all ion detectors count discrete events (ion arrival) the output should obey Poisson statistics, this results in a Gaussian noise pattern that is modulated by the signal intensity [7]: In a TDC such as that used by the QStar Elite instrument, 'spikes' corresponding to ion arrivals are triggered by the characteristic gradient change in the detector voltage that accompanies the ion arrival. Due to this triggering behaviour thermal noise can be considered negligible whilst the Poisson nature of the incoming signal is preserved[100]. As the

baseline effect is eliminated in an orthogonal instrument a noise model can be written

$$\mathbf{x} = \mathbf{s} + \mathbf{g} + \mathbf{p}\sqrt{\mathbf{s}} \tag{5.8}$$

where $\mathbf{s}$ is the true peak signal, and each entry in $\mathbf{g}$ $\mathbf{p}$ are chosen from an i.i.d. normal distributions $N(0, \beta)$ $N(0, \alpha)$ respectively. For this calculation it was assumed that any electronic noise is negligible compared to the shot noise and so $\mathbf{g} = \mathbf{0}$.

The shot noise profile was approximated by subtracting a smoothed version of $\hat{\mathbf{x}}$ (using a moving average with a window of 7 bins) from $\mathbf{x}$ over defined peaks[201] so $\mathbf{s} \approx \hat{\mathbf{x}}$ according to the rearrangement of Equation 5.8

$$\mathbf{p}\sqrt{\mathbf{s}} \approx |\mathbf{x} - \hat{\mathbf{x}}| \tag{5.9}$$

$$\mathbf{p} = \frac{\mathbf{x} - \hat{\mathbf{x}}}{\sqrt{\hat{\mathbf{x}}}} \tag{5.10}$$

This was calculated for every pixel in the image providing a noise distribution and the variance of the

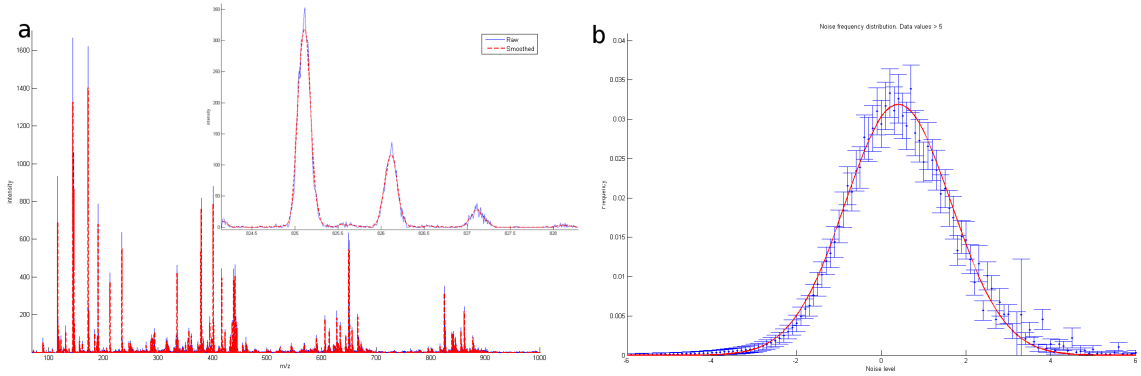

Figure 5.6: Estimating the spectral noise distribution. a) To estimate the noise level per mass bin a smoothed version of each spectrum (red dash) was subtracted from the original data(blue). b) Calculating the noise levels for every pixel in the image gives the smooth Gaussian profile expected with $\alpha = 1.25$

Gaussian function underlying the shot noise (α) was extracted by fitting a Gaussian distribution to the of values of $\mathbf{p}$ (see Figure 5.6). Only mass bins in $\mathbf{x}$ that had non-zero values were considered to reduce skewing of the distribution due to the spectral threshold that had been applied. This indicated that the distribution was very close to being zero mean, with some slight skew possibly introduced by the hardware threshold process, and α was found to be 1.25.

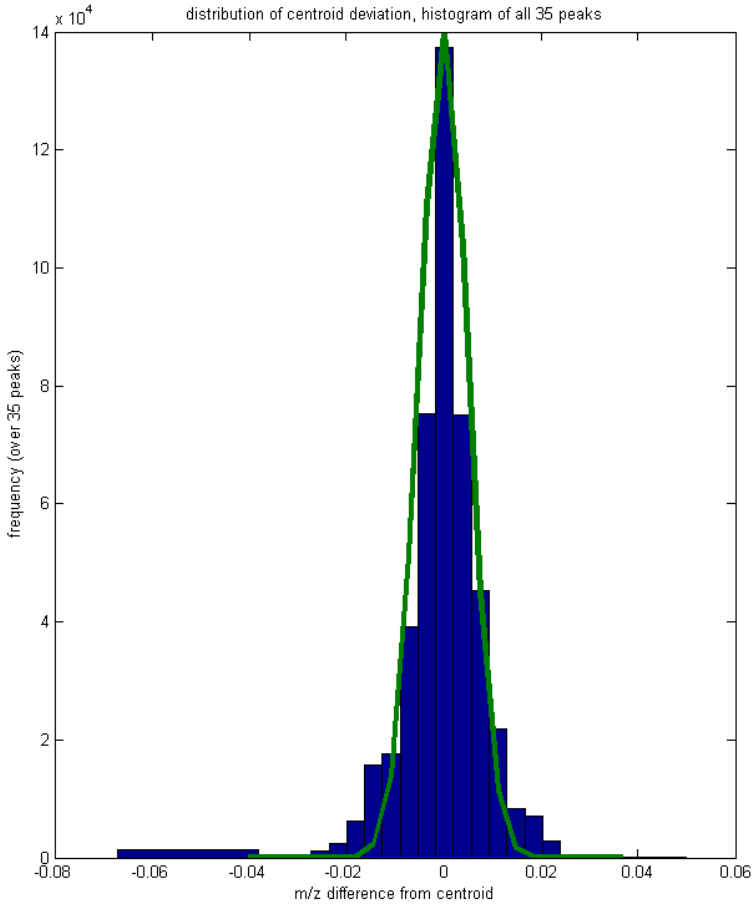

Figure 5.7: Distribution of differences between mean centroid location and peak centroid in individual spectra. Total histogram for 35 most intense peaks over 20535 spectra. Green fit shows a normal distribution with standard deviation 0.005

5.3.5 Mass Accuracy

The mass accuracy measures the difference between an expected m/z value and a centroid in the measured data $\Delta m_i = m_i - m_{expected}$. There is some spectrum-to-spectrum variation in the centroid location for a particular peak. Whilst fitting the Gaussians in Section 5.3.3 the centroid of the fitted Gaussian was also recorded and the difference between the centroid in the mean spectrum and every individual spectral was recorded. Figure 5.7 shows the distribution for the aggregate of all 35 peaks. Fitting a normal distribution to this histogram yields a standard deviation of $\delta_m = 0.005 m/z$.

QStar Elite Model Parameters

Component	Function	Parameter Value (units)
Mass analyser	Centroid Variation	$\delta_m = 0.005$ (standard deviation, m/z)
	Gaussian convolution	$\kappa = 0.1-$ (standard deviation m/z)
Mass detector	Additive Poisson Noise	$\alpha = 1.25$ (scalar)
	Quadratic Time Binning	$\Delta_t = 1.8e^{-4}$ (s)

Table 5.1: Simulation parameters established for the generation of the simulated data sets for the QStar Elite instrument

5.3.6 Validation: Reverse Engineering the Brain Image

The model detailed in Table 5.1 alongside the experimentally determined parameter values extracted from spectral data has been defined to reproduce spectra as generated by a QStar Elite instrument, this ability is tested by reverse engineering a real-world dataset. The important output of any model is an evaluation of the similarity between the output and compatible physical data, to evaluate this the fixed rat brain image was used as an example of a real-world dataset. Success can be determined by the similarity between the real data and the spectral output of the simulated data.

For simulation the dataset must be decomposed into a set of ion-lists with accompanying spatial abundances. The general approach taken here is to reduce the data to a collection of peak centroids and then factorise these to produce a small number of basis vectors that reconstruct the spatial and spectral distributions. The full dataset was reduced to a peak list (by peak detection on the mean spectrum maintaining 1000 peaks, see Chapter 4), and a datacube constructed by summing the spectral intensity within a window of ± 0.25 m/z in each spectrum. NNMF was then used to decompose the data as it produces a set of spectra that can be combined additively to reconstruct the input, the components are shown in Figure 5.8). The NNMF captured most of the spectral and spatial variation, a pixel-wise evaluation of the total difference between the datacube and a reconstruction from the NNMF factors produced an average correlation of >0.99.

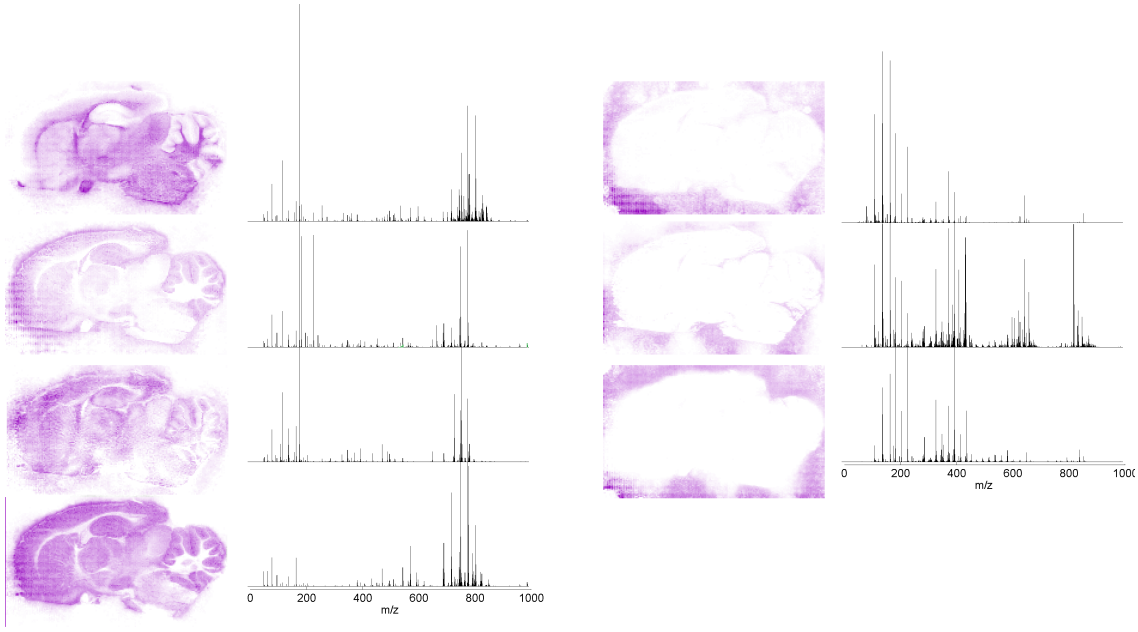

Figure 5.8: Generating inputs for simulation using NNMF. NNMF factorisation output produces a set of spectra with spatial weighting for each over the image such that the sum of all factors reconstructs closely the original data. These spectral and spatial outputs of NNMF with 7 factors are shown here.

The research question of whether the various decompositions provide insight into the molecular composition of specific tissue types is still ongoing[108] but in this case it is simply a collection of layers sought that produce molecular signals at each pixel that are representative of an actual dataset. These factors then provide an ion-list for each pixel with intensities that can be linearly combined to reconstruct the original data. They therefore make an ideal starting point for use with the simulated data for modelling the instrument they were produced on.

Comparing the Output of the simulated data with the raw data

Initial ion-lists produced from the NNMF decomposition of the fixed rat brain MSI data were processed with the simulator using the parameters shown in Table 5.1. The peak centroids from the ion list were turned into noisy peaks in the simulated output which were then compared with the raw mass spectrometry data. When viewed side-by-side, as seen in Figure 5.9, individual spectra from the original dataset and the simulated show near identical intensities over the peak width and the same noise characteristics. Pixel-to-pixel statistical variation were measured from the data so ion images also show similar noise characteristics over the image area, as illustrated in Figure 5.10. The properties can now be transferred to any set of input ions and spatial distributions.

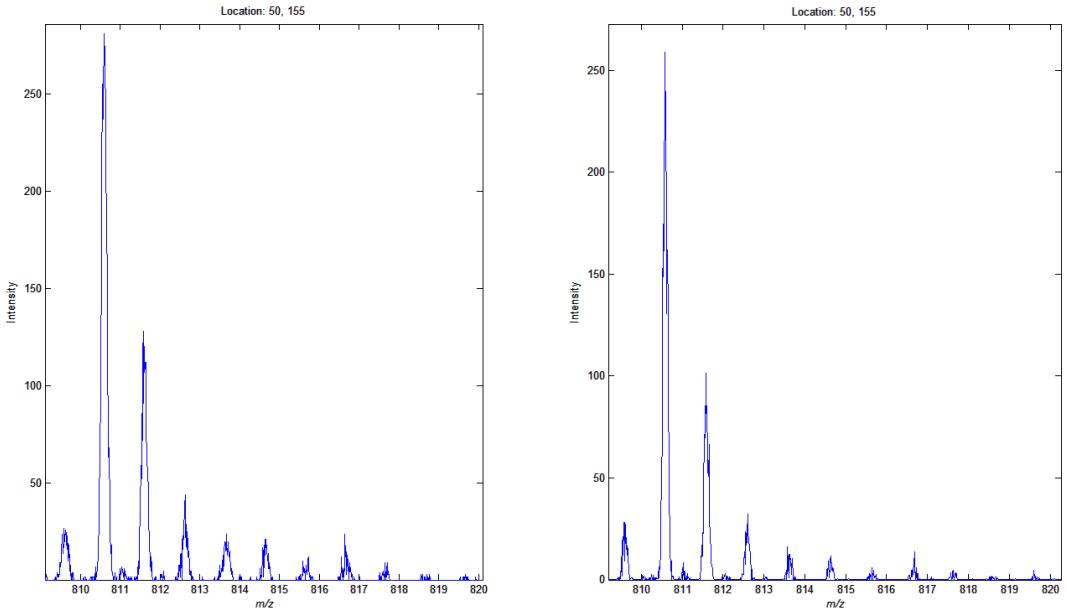

Figure 5.9: Zoom on a few peak from (left) raw data and (right) simulated data. Some small differences in intensity exist due to both the NNMF process and random factors but the peak shape and noise character have been reproduced.

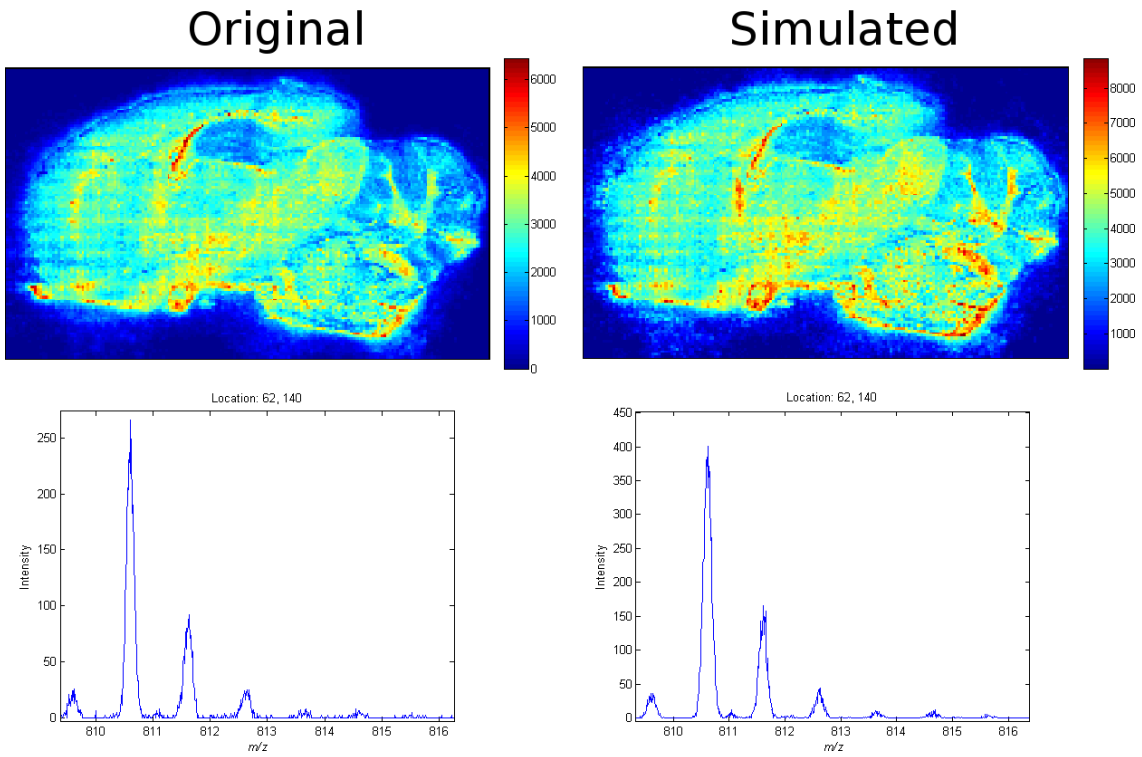

Figure 5.10: The output of the simulation is compared to the original dataset. Top row: ion image of m/z 810.6 ± 0.05 generated from a simulated data dataset and a raw MALDI image. Bottom row: spectra from an identical spatial location in the two datasets. The distribution and noise have the same characteristics in both the simulated data and the MSI dataset that was reverse-engineered.

5.4 Software Interface

A GUI has been developed which allows different detectors and mass analysers to be selected for the same input data, illustrated for a typical sequence of operations in Figure 5.11. This allows the simulation to be tuned for a specific instrument, as in this Chapter, or for a standard dataset to be simulated for several instruments. The software interface is designed to allow an overview of all currently selected options. By design, it should be straightforward to add new instrument options, providing the function is crafted according to the specification and placed in the appropriate folder it should be automatically detected and included in the simulator.The available functions are automatically detected from folders within the simulator directory, if additional functions are added these lists can be refreshed, the current combination of functions will be referred to as the instrument list.

On adding a function to the current simulator pipeline a dialogue box is called within which the user can

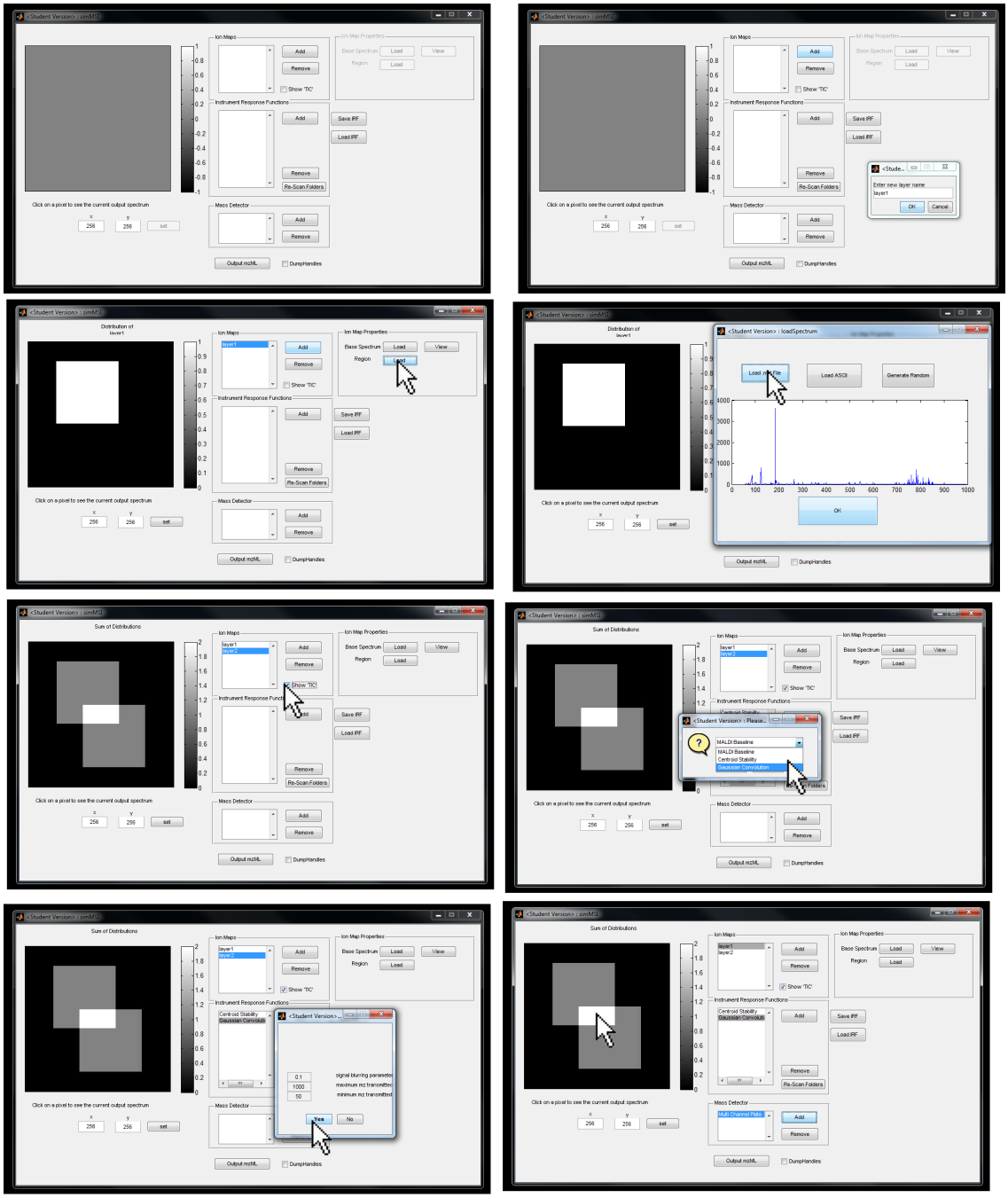

Figure 5.11: A graphical interface for producing a simulated dataset. The stages of adding a dataset are (from top row, left to right): The interface; add and name a layer. Load a spatial intensity. Set the ion list. The 'TIC' check shows the sum of all layers added. Add an instrument response function (in this case convolution). This brings up the detected options for that function. Clicking on a pixel in the preview window generates the spectrum for that pixel. These spectra show the characteristics of the modelled instrument.

input parameters for that function. The parameter list is detected by the software by querying the function, each function must return a properly formatted list of parameters, default values and a description.

The m/z range is set and the software determines the m/z domain automatically by querying the detector function (after user parameters have been set) and discretises the space according to the detector resolution. The effect of the current instrument list on a spectrum can be judged by clicking on pixels within the data map viewer, this takes the selected pixel and performs the complete spectrum generation, as shown in Figure 5.11. This route can be used to manually experiment with the effect different noise sources have. When the desired instrument list has been input the compete image can be generated automatically. The net result of these stages, illustrated in Figure 5.1 is a spectrum at each pixel with peaks that have the characteristics of the instrument being modelled.

5.4.1 Data Output

Output was produced directly in the imzML format using the parser included with imzMLConverter (v.1.08)[173]. Using an openly accessible data format provides the capability for easy file sharing and comparability across different software packages.

5.5 Simulated Data-Set

Using the MSI simulator realistic datasets with known spectral and spatial contents can be produced. This section describes one such dataset which will later be used as a ground truth for the evaluation of clustering algorithms for automated image segmentation.

Simple shapes

This simulated dataset is designed for visual clarity rather than biological-mimicry. It has four layers seeded with a different ion list, as illustrated in Figure 5.12. Three contain spatially localised shapes which are present in isolation and overlap. A background region with varying intensity is superimposed over the whole image to represent the situation in MALDI where matrix is present in all spectra. All regions have randomly generated peak lists in the m/z range 100-1000, enumerated in Table 5.2. This is a conceptually straightforward problem but also introduces issues of boundary areas where mixed profiles are present.

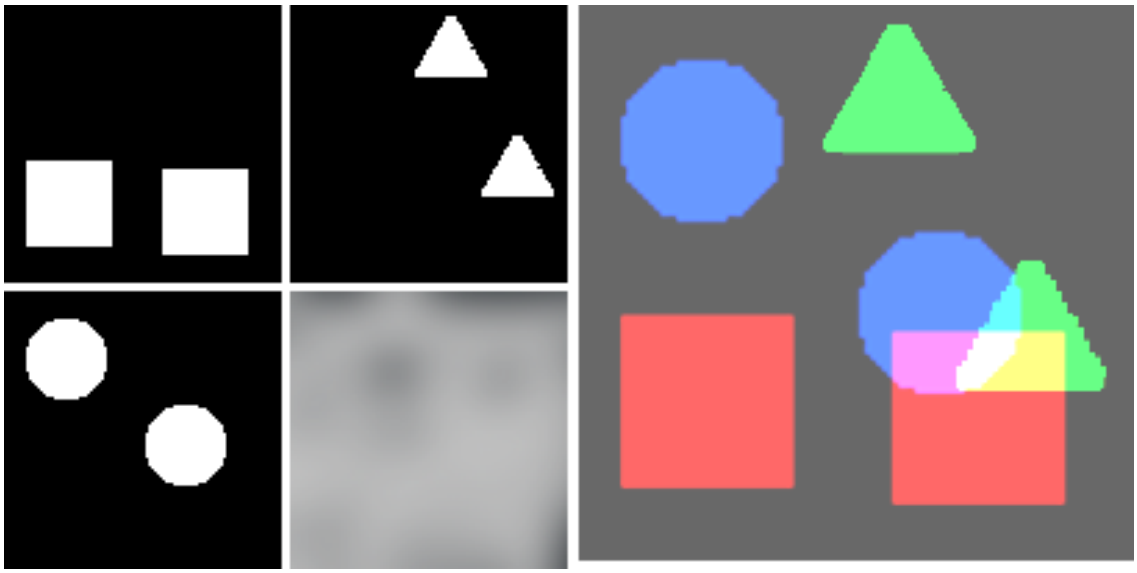

Figure 5.12: A simulated dataset consisting of simple shapes which consists of a set of overlapping geometric shapes with a fluctuating background. Each shape shares the same spectrum, mixed colours indicated overlaps and unique spectral regions.

Region	Shape	Number of Peaks
1	Squares	100
2	Triangles	100
3	Circles	100
4	Background	200

Table 5.2: Number of random peaks used as the ion list for each of the regions shown in Figure 5.12

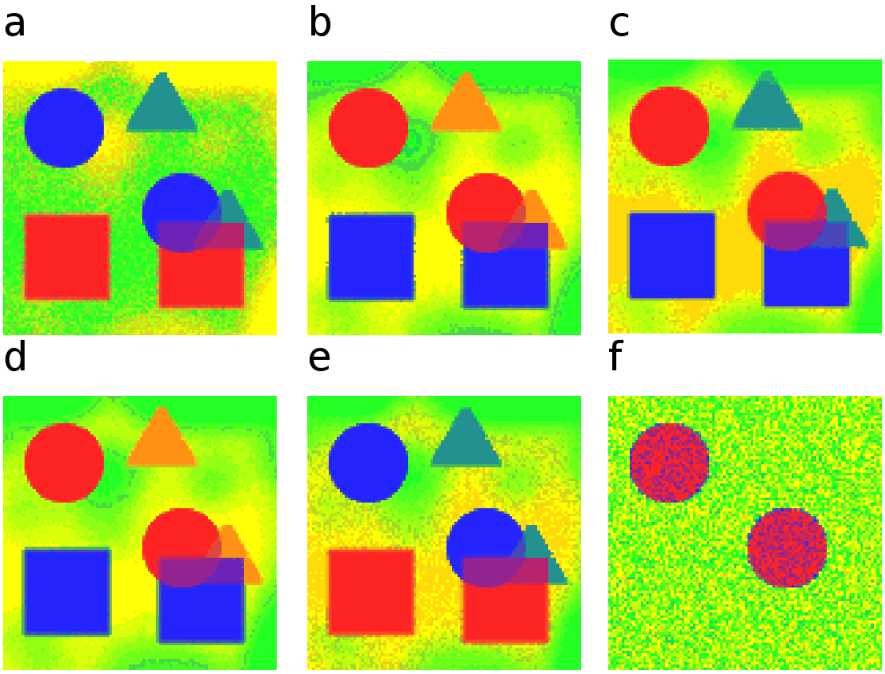

Figure 5.13: Feature selection pipelines evaluated on simulated data. The feature selection pipelines from Chapter 4 were applied to the simulated dataset and a SOM used to evaluate the known features that were retained (see Chapter 4 for details on the pipelines: A. none; B. Re-bin 0.2; C. STD Spectrum; D. Summary Spectra; E. Frequent Peaks; F. Spatial Correlation).

5.5.1 Feature Selection Evaluation with Simulated Data

The development of a data simulator was motivated by the lack a of ground truth for analysing MSI data processing algorithms. To illustrate how simulated data may help with this task the processing pipelines detailed in Chapter 4 were applied to the 'shapes' simulated data. Again, visualisation using a SOM following Principal Component Analysis (PCA) was used to evaluate the output of each pipeline. The resulting classification maps are shown in Figure 5.13.

It is interesting to note that variations in background intensity are picked up as distinct clusters in all classification maps apart from the spatial correlation pipeline (Figure 5.13.f) where it is completely removed. This is informative, as none of the pipelines attempt to correct for systematic intensity fluctuations with a normalisation step. Figure 5.14 shows a SOM classification map of Basis Approximation for Spectral Compression (BASC) compressed data that has been subjected to Total Ion Chromatogram (TIC) normalisation and the effective removal of the background variation.

Also notable in Figure 5.13.f is that two groups of shapes are missing, the square and triangular regions. All of the ion images corresponding to these shapes are removed as uninformative as the threshold for

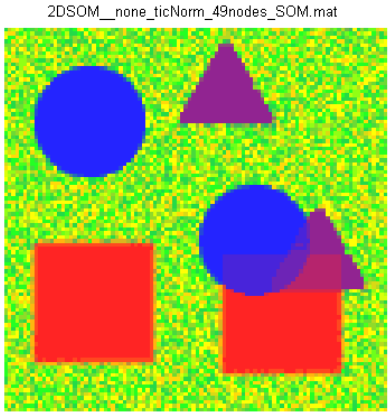

Figure 5.14: Applying TIC normalisation of BASC compressed data prior to visualisation with a SOM removes the effect of background intenisty fluctuations.

correlation with a background signal is not set appropriately for this dataset (the algorithm assumes that non-background signals anti-correlate with background ions). This highlights the requirement of determining suitable algorithm thresholds for the type of image scene encountered, a task which is explored for spectral clustering in Section 5.6.

There is a small region of overlap (approximately 50 pixels or 5% of the image) formed at the intersection of all three shapes (coloured white in Figure 5.12). This was not visualised following any of the peak detection methods but it is identified from both the direct BASC and BASC following re-binning pipelines. Without having these baseline methods to compare against, it would not have been obvious that such a small area was omitted in the other classification maps and it would be worthwhile to investigate the cause of the discrepancy between methods.

5.6 Segmentation with Spectral Clustering

To illustrate the utility of a simulated dataset it is used to tune the parameters of a popular segmentation algorithm that has not yet been applied in MSI (See Section 1.4.5). It has been shown to be a powerful approach for clustering high-dimensional data [220] and here it is presented for segmenting mass spectrometry data. Spectral clustering differs from traditional clustering algorithms, like k-means, in that the clustering is not performed directly on the similarity metrics between pixels. Instead, a graph Laplacian is formed and clustering is performed on its eigenvectors to determine which nodes on the graph (datapoints) are substantially connected (have a relatively large similarity). To perform successfully the similarity graph

must be sparse so that overall connectivity is only high between members of the same cluster. This is often achieved using a nearest-neighbours method, so that only values for a user specified number of most similar points are populated in the similarity matrix.

Algorithm 5.1: Spectral Clustering

Data: Data matrix $\mathbf{A}$, number of nearest neighbours k, distance parameter σ

Result: cluster membership vector $\mathbf{c}$

for *i=1:number of spectra* **do**

1 Find the the k nearest neighbours of $\mathbf{A}_i$ **for** *j=1:k* **do**

2 Calculate a similarity matrix for pairs of spectra $\mathbf{A}_i$ and $\mathbf{A}_j$ using the Gaussian distance $\mathbf{S}_{ij} = e^{(\frac{-||\mathbf{A}_i - \mathbf{A}_j||_2^2}{2\sigma^2})}$;

end

end

3 Construct the degree matrix $\mathbf{D}$, a diagonal matrix where each element of the diagonal containing the total connectivity of that data-point $d_i = \sum_{j=1}^{n} s_{ij}$;

4 Calculate the graph Laplacian $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{S}$ (where $\mathbf{I}$ is the identity matrix);

5 Compute the eigenvalue decomposition of $\mathbf{L}$;

6 Partition the data recording the cluster membership into a vector $\mathbf{c}$ either with k-means;

5.6.1 Spectral Clustering of Simulated Data

The spectral clustering toolbox implemented by *Chen et al*[37] was used with modification to use the random walk Laplacian $\mathbf{L}_{rw}$. In this implementation an undirected similarity graph is produced as an edge is added between any points that are nearest neighbours. Note that exponent of the Gaussian similarity metric is the Euclidean distance which can be calculated from data compressed using the basis approximation algorithm.

The σ term in Algorithm 5.1 (see also Equation 1.13) is a parameter that controls the neighbourhood size and was decided using the self-tuning method[37]. As most values of the similarity will be zero, it is not computationally efficient to calculate each pairwise similarity. Instead only the similarity to the nearest k neighbours for each spectrum will be calculated. The sensitivity of the clustering to the number of nearest neighbours will be explored on the simulated dataset.

The 'shapes' dataset presented in Section 5.5 was used as the ground truth for this evaluation and a simulated dataset was produced using the 'QStar Elite' instrument model shown in Table 5.1. For evaluation of the segmentation the ground truth were specified as the regions within the dataset where the input maps

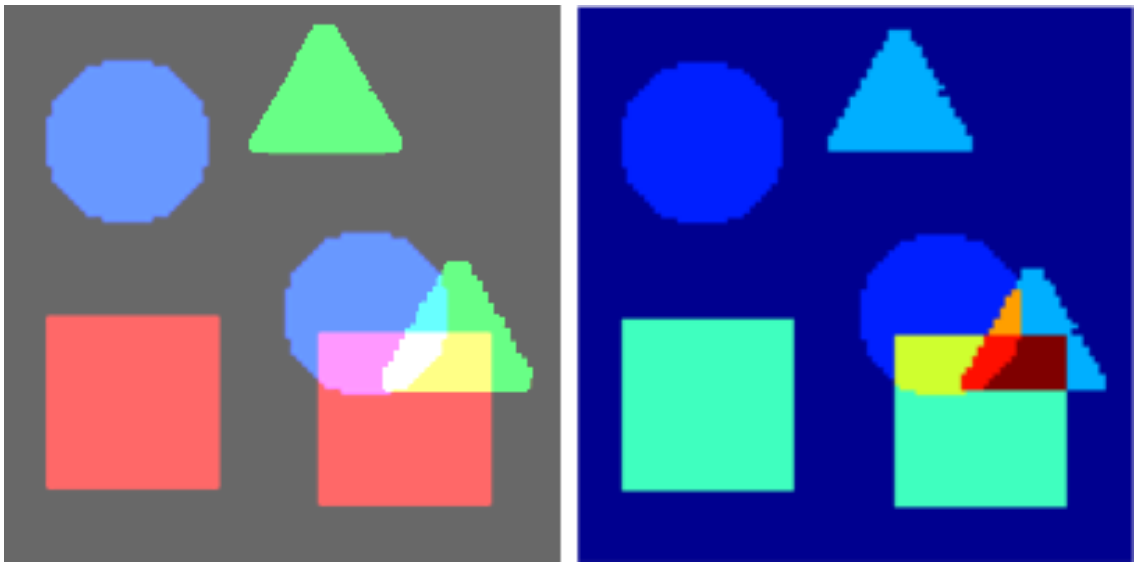

Figure 5.15: Dataset for evaluating segmentation using spectral clustering. The 'shapes' simulated dataset was used. The left image illustrates the four unique spectral inputs and where they overlap. This produces 8 distinct regions of spectral similarity which are colour coded in the image on the right.

overlapped. As is illustrated in Figure 5.15 these are the areas of background and shapes (3); background and shape overlaps (4); and background only (1) to provide a total of 8 unique spectral profiles. After simulation the data was dimensionality reduced using 200 random projections without further processing.

Evaluation metrics

In a binary classification scheme, algorithm effectiveness can be quantified in terms of the number of pixels which are correctly identified: True positive (TP); number of pixels incorrectly identified: False Positive (FP); number of pixels correctly rejected: True Negative (TN); and the number of pixels incorrectly rejected: False Negative (FN). From these several performance measurements can be calculated: accuracy (ACC) measures the rate of correct pixel classification

$$\text{ACC} = \frac{TP + TN}{TN + FP + TP + FN} \tag{5.11}$$

the negative predictive value (NPV) gives the rate at which there pixels are appropriately classified as not being a cluster member,

$$\text{NPV} = \frac{TN}{TN + FN} \tag{5.12}$$

the positive predictive value (PPV) measures the rate of appropriately classified as being a cluster member,

$$\text{PPV} = \frac{TP}{FP + TP} \tag{5.13}$$

sensitivity (SE) measures the method's ability to detect a cluster member in the data,

$$\text{SE} = \frac{TP}{TP + FN} \tag{5.14}$$

specificity (SP) measures the method's ability to correctly identify the absence of a cluster member in the data,

$$\text{SP} = \frac{TN}{TN + FP} \tag{5.15}$$

In the multiclass case these are calculated for each class against the combined members of other classes. Figure 5.16 shows the regions of TP,FP,TN,FN within a multi-class classification array for one cluster.

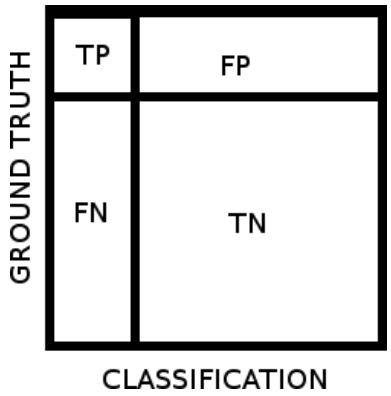

Figure 5.16: Elements of the confusion matrix defined for label 1 within a multi-class tabulation. a_{ij} corresponds to the number of pixels with true class i labelled as j.

5.6.2 Evaluating the number of nearest neighbours.

Spectral clustering takes two inputs, an expected number of clusters and the number of nearest neighbours to consider when building the similarity graph, kNN. The kNN parameter influences the sparsity of the graph Laplacian and considering too many neighbours may artificially connect pixels that are from disparate regions, negatively impacting the clustering result. There has not yet been a principled method for choosing this value [220] so it is proposed that a simulated dataset allows for the estimation this parameter for subsequent clustering.

The 'shapes' simulated dataset (100×100 pixels, 114311 m/z channels) underwent dimensionality reduction by random projection (200 projections), no other data process was performed. Spectral clustering was performed using 8 clusters and a range of values for kNN from 1-20. The spectral clustering is initialised differently each time and so the image labels do not have a pre-defined order. In order to compare the results with the ground truth shown in Figure 5.15 the region labelling was manually aligned (without altering the pixel connectivity). In each case the multi-class confusion table was produced and from this, the accuracy was calculated.

Figure 5.17 shows the classification accuracy plotted against the number of nearest neighbours. This plot shows that nearest neighbour values from 3-8 provide optimum clustering, and for this range a clustering accuracy of greater than 90% was achieved. In contrast to a density based method such as the SOM (Figure 5.13) the segmentation is not dominated by the number of pixels in the background. When the number of nearest neighbours exceeded 8 the smallest cluster (orange in Figure 5.15) was lost and intensity variations in the background start to be divided into separate clusters. The background subdivisions correspond to the intensity variations present there, see Figure 5.12, this is not surprising as the Euclidean distance amplifies intensity differences and suggests that intensity normalisation may be appropriate to ensure that spectral clustering operating only on relative spectral changes.

5.6.3 Effect of number of random projections on clustering

The simulated data can be considered to both optimise the performance of spectral clustering in absolute terms and evaluate it relative to another common clustering algorithm, kmeans. The 'shapes' dataset underwent dimensionality reduction using random projection where the number of projections used was varied from 1 to 200. Each set of projected data was then clustered using both the spectral clustering and k-means clustering algorithms. Both algorithms require a pre-specified number of clusters and 8 clusters were specified to each (the expected number of regions in the 'shapes' dataset). Spectral clustering used 6 nearest neighbours (see Figure 5.17) and k-means took the best of 5 replicates. Receiver Operating Characteristics (ROC) allow an assessment of both the specificity (how often an algorithm classifies 'true' points correctly) against the specificity (how often an algorithm classifies 'false' points correctly) of an algorithm. ROC values are usually visualised by plotting the sensitivity against 1-specificity to form a ROC curve, with the best response possible being at [0,1] in the top left corner. The sensitivity and specificity was calculated independently for each region in Figure 5.12 and averaged to provide a final value for each number of projections.

A side-by-side comparison of the sensitivity to the number of random projections shown by spectral

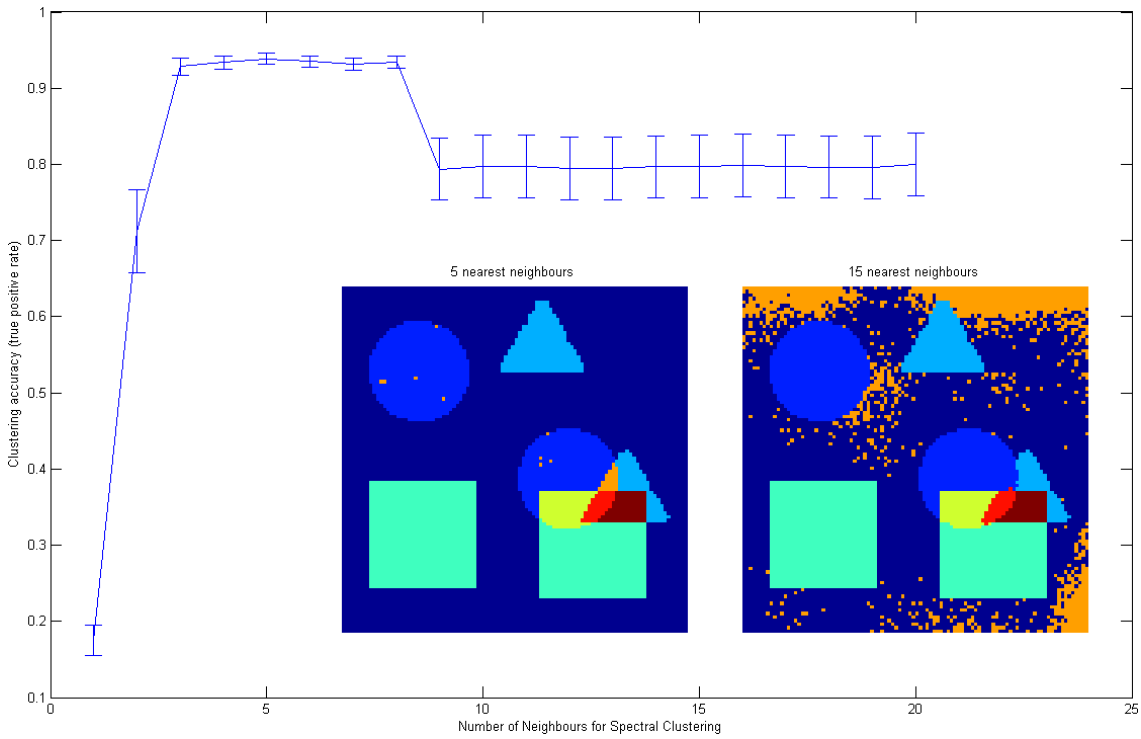

Figure 5.17: Optimising the number of nearest neighbours to use for spectral clustering. The blue line shows the true positive rate, which peaks when 3-8 nearest-neighbours are used. Greater than this and intensity variations in the background overwhelm one of the cluster regions. The inlaid regions show kNN = 5 and kNN = 15, illustrating the previous observations.

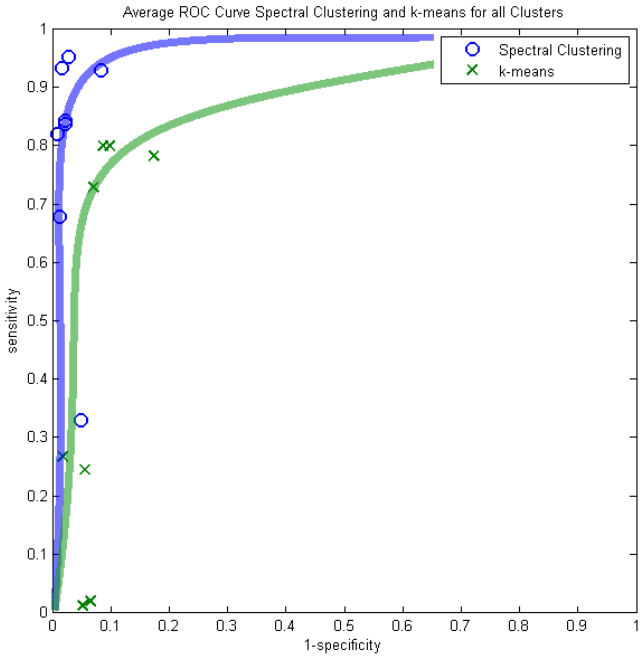

Figure 5.18: ROC curve for k-means clustering and spectral clustering. Each marker shows the average sensitivity and specificity for the image regions for at different number of random projections. The open circles are spectral clustering and the crosses are k-means clustering. Lines are produced manually to provide a visual guide only.

clustering and k-means clustering was performed. Some comments on a suitable number of projections were made in Chapter 2 but this simulated dataset allows a quantitative evaluation for the specific purpose of segmentation as well as establishing the relative merits of the two clustering algorithms. Figure 5.18 show the ROC curves for spectral clustering and kmeans with different numbers of random projections. The performance of both algorithms improved with a greater number of projections (data not in the graph, but the markers with higher sensitivity are from greater number of projections). Overall the spectral clustering ROC curve sits above the k-means clustering indicating that spectral clustering outperforms k-means. This result may reflect the underlying assumptions that each method makes about the data. Applying the k-means algorithm to detect compact clusters assumes that each cluster has approximately equal number of member pixels which are normally distributed within the data space whilst the connectivity based spectral clustering assumes a minimum level of difference between clusters. In this simulated dataset a significant difference in the size of clusters exists and this may contribute to the relatively weaker performance of the k-means algorithm. This discrepancy between cluster sizes is likely to hold true for mass spectrometry imaging in general, and so connectivity based methods may provide for a more robust segmentation.

5.7 Conclusions

A simulation tool has been developed for the evaluation of mass spectrometry imaging data processing against a known ground truth to provide a quantitative measure of algorithm performance. The IRF model presented here has been designed to reproduce data with noise characteristics of a specific instrument. Once an instrument is characterised its IRF can be easily shared so that data of that type can be assessed by other algorithms in the community. An added advantage of a simulator is that only the parameters and input distributions need to be shared for the data to be reproduced, rather than transferring large MSI datasets. The simulator has been designed with extensibility in mind if more sophisticated instrument models are required they may be rapidly deployed and evaluated on the same data-sets.

Ideally a controlled substrate with a number of known inclusions would be used to produce the instrument characteristics but it is shown here that a good reproduction of the spectra can be made from a real tissue image. As models of ion/analyte interactions become more sophisticated the ability to predict the ratios of matrix clusters to internal molecules will improve, outputs from these models can be used to seed the ion layers whose progress is then modelled through the instrument. For biological interpretation of real-world samples it should be remembered that the signal is a cumulative product of sample, ionisation and mass analysis, not just the underlying tissue.

Data analysis algorithms have been shown to accurately uncover patterns in the detected signals. To understand the performance of algorithms only an initial ion list is required combined with a physical model of the mass analysis process. With these elements algorithms can be evaluated for sensitivity to instrument parameters and variation in the initial ion list. The spectral clustering algorithm was found to have better sensitivity and specificity than the kmeans algorithm and so should be studied further for application in the interpretation of MSI data.

Chapter 6

Applications to Biological Samples

In this chapter the newly developed techniques are applied for unsupervised analysis of biological samples. The transformation of the data to a randomised basis is utilised for accelerated dimensionality reduction and to separate and concentrate spatial and spectral patterns. Matrix Assisted Laser Desorption Ionisation (MALDI)-MSI of whole porcine eye and multiple sections of human liver are explored to identify areas with distinct molecular profiles.

6.1 Introduction

Early work with MALDI-MSI focussed on the analysis of intact peptides and proteins with larger masses as ions produced from the matrix itself are very abundant in the low mass range (m/z<1500) and overlap with peaks from endogenous small molecules. Advances in instrumentation have allowed for imaging experiments to collect spectra with sufficient resolution to separate these peaks and opened this mass region to imaging experiments[39]. Without tissue washing stages the majority of ions detected are from lipids[158].

Lipids are an important class of biological molecules which are involved in many aspects of cellular activity. They are the major constituent of the cell wall as well as being important mediators in cell signalling and metabolism[158]. They have become the subject of increasing volumes of research within mass spectrometry

imaging due to the role they play in all aspects of cell processes[74].

Intensity patterns from individual lipids have been shown to discriminate between tissue types and disease states[56, 164] patterns of relative abundances of groups of lipids and other low molecular weight species (molecular profiles) have proved to be both successful and reliable for classification of biological samples. Particularly noteworthy applications for individual spectra are bacterial species identification (using MALDI [96]) and direct tissue classification (using rapid evaporative ionisation MS[12]). However, these tools are trained on a database of known samples and then classify subsequent spectra based on the similarity to existing entries and so require substantial input from domain experts on well characterised samples in order to populate the database.

In this chapter the discriminatory properties of low mass molecular profiles are detected automatically and used to segment MALDI MSI datasets acquired from whole porcine eye and diseased human liver samples. A Self Organising Map (SOM) is used to explore the molecular landscape of the porcine eye whilst k-means clustering is used to extract the tissue compartments from an image of liver suffering from Non-alcoholic Steatohepatitis (NASH).

6.2 Porcine Ocular Tissue

Tissue samples analysed by MALDI MSI can range from single cells[130] to whole animal body sections[115] but the difficulty in producing a MALDI-ready sample gets more difficult at both size extremes. This study originated as a study of sample preparation methods for whole porcine eyes that allows thin sections through the complete organ to be collected and analysed by MALDI MSI[164]. Large aqueous organs such as an eye presents a significant sample preparation challenge. Organs from smaller animals, such as rodents[90] or reptiles[180], have been processed by simply freezing the whole tissue e.g. in liquid-nitrogen-cooled isopentane, and then sectioned whilst frozen. The key difficulty in snap freezing and sectioning large mammalian eyes stems from the inhomogeneity of tissue types as well as extreme variation in water content (>99% in the vitreous[106] compared to $\approx$ 70% in muscle tissue and 60% in the lens[92].

Changes in lipid composition or distribution are associated with multiple retinal disorders such as e.g. cataracts[11] and diabetes[167]. Within the lens, changes in lipid composition have been observed by MALDI-MSI as symptoms of the ageing process[51]. As they are important diagnostic molecules it necessary to develop methods which increase the knowledge of ocular lipid distributions[167]. It was shown by *Palmer et al*[164] that formalin fixation followed by sucrose dehydration improves the tissue resilience during sectioning so that intact sections from the whole organ could be obtained.

6.2.1 Sample preparation

Porcine eyes were obtained from animals destined for human consumption. Enucleation occurred within an hour of death and eyes were transported and stored at 4°C. Eyes were immersed in formalin and cryo-protectant solution (10% formalin, 60% sucrose mixed 1:1 v/v) for 6 hours and snap frozen in solid CO_2 cooled isopentane. Frozen eyes were stored at -80°C until sectioning. Whole eyes were attached to cryostat chucks via frozen water and cryo-sections were collected at 10µm then thaw mounted onto steel imaging plates before being dried at room temperature for 30 minutes. For the complete experimental methodology see[164].

Cryo-protectant and formalin fixation were required for intact sagittal sections to be collected from an entire eye. These were then subjected to MSI in order to measure lipid distributions throughout the whole organ. Figure 6.1 shows an example section after coating with CHCA matrix and after image acquisition with the major anatomical regions labelled. This section was acquired in the coronal plane through the eye but was slightly off-axis from the pupil. Consequently whilst the cornea (1) and the lens (2) are present in the section the pupil aperture is not. Some shrinkage of the vitreous (3) occurred during drying which had the unfortunate effect of pulling the retina and associate inner-membrane (4) away from the sclera and surrounding musculature (5). Regardless, all of the major components of the eye are visible and can be imaged in a single experiment.

6.2.2 MALDI Mass Spectrometry Imaging

This mass spectrometry dataset has previously been published for the development of sample handling practises for whole mammalian eyes[164] and a full imaging methodology can be found in that paper. CHCA matrix (5 $mgml^{-1}$ in methanol/chloroform (1:1 ratio) 0.1% TFA) was applied using an airbrush so that a visibly even coating was achieved (30 passes) whilst ensuring that the tissue was never allowed to become noticeably wet. MALDI mass spectrometry images were acquired on a QqTOF mass spectrometer (QStar XL, ABSciex, Warrington UK) equipped with a Nd:YVO_4 (355 nm, 5kHZ, Elforlight: SPOT-10-100-355, Elforlight, Daventry, UK) fibre delivered laser (50µm core diameter). Spectre were collected in positive ion mode with an m/z range 100-1000. Images were acquired in raster mode with a pixel diameter of 225 x 225 µm (0.879 s accumulation per pixel). It is known that formalin fixation is compatible with lipid analysis by MALDI MSI[32, 82] and this study also successfully detected multiple lipid species. The analysis of lipid distributions across the different tissue compartments of an intact eye provides a more comprehensive snapshot compared to approaches which were previously used, such as excising sub-portions of the organ.

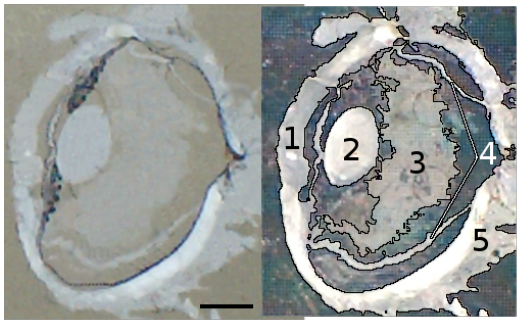

Figure 6.1: Schematic of a sagittal section through a porcine eye. a) Photographs of a porcine eye section after coating with CHCA before and after MALDI-MSI (left and right respectively). The post-analysis image is manually labelled: 1, Cornea. 2, Lens. 3, Vitreous humour. 4, Retina. 5, Sclera. There was a noticeable contraction of vitreous (3) that occurred following thaw mounting which caused some dislocation of the retina(4) (scale bar 4 mm)

Data Conversion

Mass spectrometry images were extracted from the proprietary instrument format (.wiff) to the imzML format (converting to mzML using AB SCIEX MS Converter (version beta 1.1, ABSciex, Warrington UK) then to imzML using imzMLConverter (v1.0, www.imzMLConverter.co.uk[173])). The imzML parser included with imzMLconverter was used to load individual spectra into MATLAB[173].

6.2.3 Basis Approximation

The basis approximation method (Algorithm 2.4) was used to produce an orthonormal basis from the spectra onto which the whole imaging dataset was projected yielding a low rank representation of the data. Sampling was performed with $k = 150$ random projections drawn from a Gaussian distribution (all spectra sampled). The number of samplings was chosen based on the conclusions from Chapter 2.

Compression reduced the number of spectral channels from 129,750 to 150 and so provided a reduction in the overall dataset size, originally $\mathbf{X}$ with $m \times n = 115250 \times 6975 = 803868750$ elements to the basis-projection pair $\mathbf{Q}$ and $\mathbf{A}$ with elements $k(m + n) = 150 \times (115250 + 6975) = 18333750$ elements giving a compression ratio $R_c = 0.02$. The raw data size was ≈3 GB, this was reduced to ≈70 MB with this method. The SNR and PCC were calculated between corresponding raw and decompressed spectra and averaged over the dataset, giving a PCC of >0.9 and an SNR of ≈59.

6.2.4 Visualisation with a Self Organising Map

A two-dimensional SOM (as described in Chapter 3) with 49 nodes (7×7) was trained on the spatial portion of the compressed imaging dataset and subsequently used to segment the image. Computing the node mapping took less than one minute when applied to the compressed data. An empirical timing study was conduced where the SOM was trained on the full spectra, and the estimated completion time was approximately 14 hours. As has been shown within this thesis, equivalent (or improved) segmentation results are produced by operating on the compressed data the segmentation on the full data was not produced.

Each spectrum was classified according to its most similar SOM node (based on Euclidean distance of the projected data) and a RGB colour-scheme was applied to produce a segmentation map. This scheme, shown in Figure 6.2, uses colours from the 'psychological primary colours'[48], which are highly separable to humans according to the antagonistic theory of vision[102]. Each corner of the SOM is allocated a 'pure' colour and these are linearly interpolated to shade the remaining nodes. The subsequent segmentation can be seen in Figure 6.2 where the anatomical features visible in Figure 6.1 are clearly reproduced.

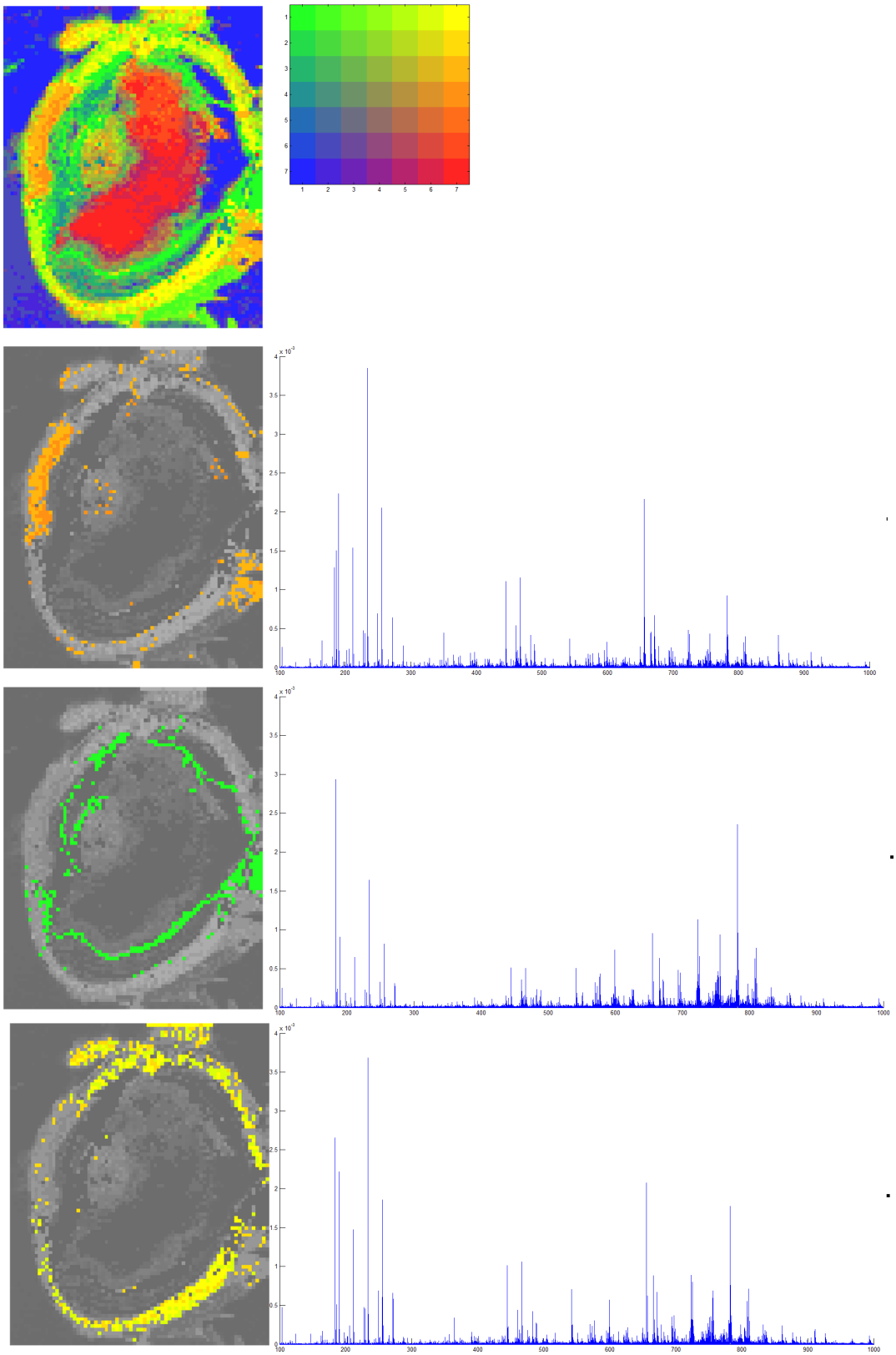

Figure 6.2: The MALDI image of a porcine eye was visualised as a single segmentation map by training a 2D-SOM (49 nodes). The resulting segmentation reveals the major anatomy identified in Figure 6.1. Several cluster centroids are also shown revealing the molecular profiles of the anatomical regions.

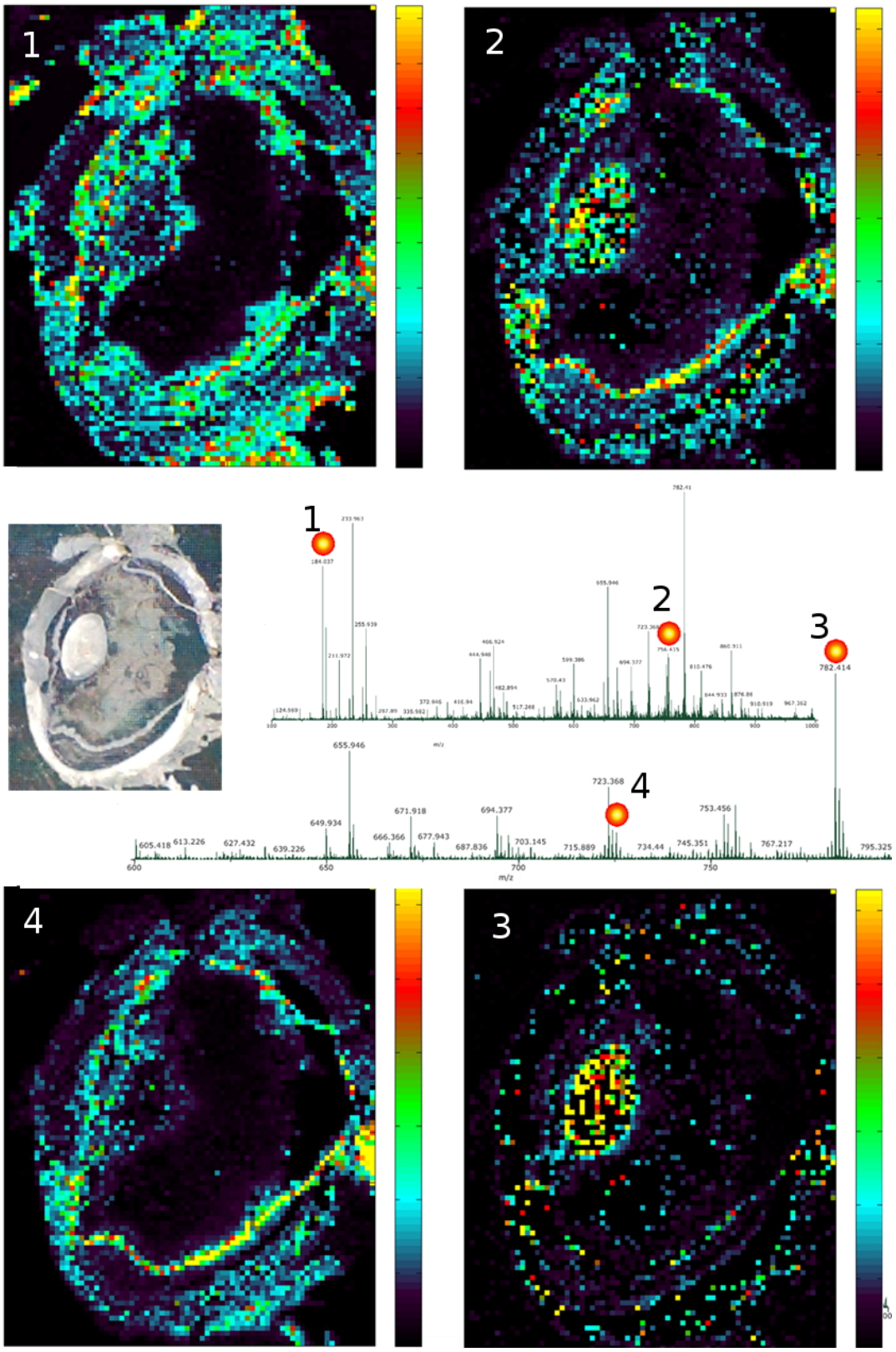

Figure 6.3: Several examples of molecular distributions observed within whole porcine eye sagittal sections. 1, m/z 184.1 Phosphocholine head group. 2, m/z 756.6 PC(32:0). 3, m/z 728.5 PC(30:0). 4, m/z 782.4 PC(34:1). Intensity limits: zero to maximum normalised ion count, per image. Example spectra from an area of retina with image m/z values indicated.

As each pixel is allocated to the node in the SOM that has weightings that are most similar to its own spectra, molecular profiles can be produced by identifying the nodes corresponding to anatomical features (by colour) and back-projecting its weightings to the m/z domain (as described in Chapter 3). Figure 6.2 shows the node weightings from SOM components that segment the lens, the retina and the sclera. From the peaks that are notably more intense in the weightings it was possible to determine multiple lipid maps that highlighted specific areas of anatomy and the identities of some of these ions were confirmed with MS/MS mass spectrometry. A selection of these are shown in Figure 6.1 where m/z 184.1 is highly likely to be the phosphocholine head group which is a fragmentation product of Phosphatidylcholine (PC) lipids. As these lipids are integral parts of cell membranes they are ubiquitous across cell types so this fragment highlights all the tissue areas. Further lipids with more localised distributions were identified by MS/MS[164] at m/z 756.6 was PC(32:0) and showed equally throughout the lens and retina; m/z 782.4 was PC(34:1) was detected within the retinal layer but at low abundance within the lens; and m/z 728.5 was PC(30:0) and was primarily detected within the lens. All these lipids were confirmed as being sodium ions ($[M+Na]^+$) in line with findings of other work examining lipids in formalin fixed tissue[32]. No ions were detected that absolutely discriminated between tissue types, only intensity differences were detected between the SOM node weightings.

After methods were developed for handling large aqueous samples it was possible to perform MALDI MSI experiments. In contrast to other work which has excised specific ocular tissue sub-regions this work presents ion images showing the distribution and localisation of certain lipids throughout the porcine eye.

The task of identifying small molecule profiles that characterised and differentiated specific tissue regions was accomplished using segmentation with a self-organising-map. The data analysis is totally automated up to the point at which the colour-coded classification map is produced at which point a biological domain expert could identify anatomical regions simply by colour and produce a molecular profile from the SOM weightings. Even though the SOM is calculated on the compressed data the molecular profiles can be produced in the m/z domain. Further analytical work to confirm the identify of key molecules could then be guided by the images, which show which tissue areas featured high levels of detection of a particular species.

6.3 Diseased Human Liver

NASH disease is characterised by the accumulation of fat within liver hepatocytes (steatosis) and in a proportion of patients this is followed by the development of necroinflammatory activity that leads to cirrhosis[64, 120]. The development of liver cell ballooning and inflammation (steatohepatitis) determines

whether a patient progresses to irreversible liver damage and fibrosis[124]. As the causes are not well understood often the only medical treatment option available is a transplant so accurate diagnosis of the condition is essential and can currently only be achieved by histological examination[9].

6.3.1 MALDI MSI of human liver

This mass spectrometry dataset has previously been used for the illustration of mass spectrometry image visualisation[67] and a full imaging methodology can be found in the supporting information of that paper.

Tissue Handling Samples were collected from patients undergoing liver transplantation or tumour resection surgery at The Queen Elizabeth Hospital in Birmingham, with local research ethics committee approval (NHS Walsall LREC) and written informed patient consent during transplantation surgery. All samples were rapidly processed and snap frozen in liquid nitrogen prior to storage at -80 °C.

Sectioning Serial tissue cryo-sections of NASH liver were obtained at 5 µm and collected either onto steel MALDI target plates (ABSciex, Warrington, UK) for mass spectrometry or glass slides destined for routine Haematoxylin and Eosinn (H&E) staining and optical photo-microscopy.

MALDI Imaging α-cyano4-hydroxycininnamic acid (CHCA) (15 mg mL^{-1} in 80% CH_3OH, 0.1% TFA) was applied to the sample and MALDI plate using an artist airbrush. MALDI TOF MS analysis was carried out on a hybrid quadrupole time of flight mass spectrometer (QStar XL, ABSciex, Warrington UK) equipped with a Nd:YVO_4 (355 nm, 5kHZ, Elforlight: SPOT-10-100-355, Elforlight, Daventry, UK) fibre delivered (100µm core diameter) diode pumped solid state laser, providing a mass resolving power of >6000 at m/z 643. Spectra were acquired in positive ion mode with a spatial resolution of 100 µm in both x and y directions.

MSI datasets were collected from serial sections of a liver suffering from NASH and from a healthy liver. The pathological differences in the liver cells is visible in ion images of m/z 798.5 (as confirmed by histopathological examination) are shown in FIgure 6.4 and illustrate the homogeneous appearance of normal liver as opposed to the enlarged and inflamed hepatocytes within the NASH diseased tissue.

It was investigated whether MALDI-MSI was able to automatically identify the different tissue regions seen within NASH affected liver. If possible this could provide support for diagnostic classification of NASH, and aid in understanding the molecular characteristics of the disease. It was considered too labour intensive to manually investigate every ion image within the complex spectra produced from this tissue, seen in Figure 6.4, so a workflow of compression using BASC followed by spatial segmentation was performed. This successfully

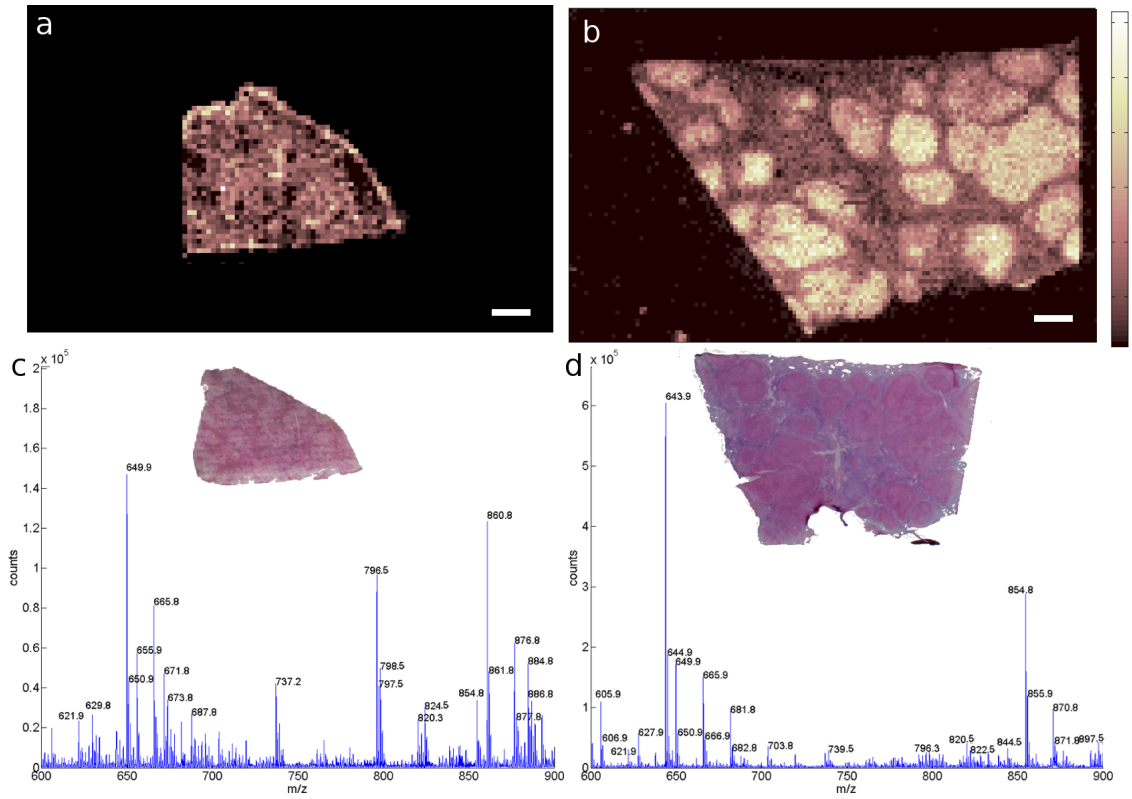

Figure 6.4: a) Ion images of m/z 796.5 $\pm$ 0.05 from healthy tissue b) Ion images of m/z 796.5 $\pm$ 0.05 from diseased tissue c) Mean spectrum of healthy tissue MSI dataset with H&E micrograph inset d) Mean spectrum of diseased tissue MSI dataset with H&E micrograph inset. Schematic of the liver image showing characteristic histology of NASH disease including fibrotic tissue (pale) and enlarged hepatocytes (dark) with H&E macroscopy inset. Ion image of showing greater intensity in parenchymal areas of hepatocytes separated by bands of fibrotic tissue with much lower signals (scale bar 1 mm) total spectrum from NASH tissue

detected the histologically known tissue regions and allowed a set of molecular profiles to be produced for each by averaging all the spectra within the region.

6.3.2 Representation with BASC

The MSI datasets fro the diseased liver sample were independently compressed by generating a BASC model for each (Algorithm 2.4) using 150 samplings. The comparative sizes are shown in Table 6.1

Section Index	Original Image bytes $[m \times n]$ (Size GB)	BASC Model bytes $[k \times (m+n)]$ (Size GB)	Compression Ratio
1	33725×12325 (3.1)	$150 \times (33725 + 12325)$ (0.05)	0.016
2	33725×10773 (2.7)	$150 \times (33725 + 10773)$ (0.05)	0.018
3	33725×12255 (3.1)	$150 \times (33725 + 12255)$ (0.05)	0.016

Table 6.1: Imaging dataset dimensions, size and compression ratios. Each image from serial tissue sections (numbered by sequential index) was collected with the same mass axis (length m) and contained n spectra. An individual BASC model was constructed for each using the same number of samplings k. The calculation of the number of bytes required to store the data in each case is shown and the disc storage size calculated. This does not take into account the storage of any meta-data, purely the spectrum intensity values.

Segmentation with Spectral Clustering

Segmentation was performed on slice 1 using spectral clustering, as introduced for MSI in Chapter 5, using 7 clusters, as shown in Figure 6.5. The cellular changes caused by NASH effect individual cells and cause them to enlarge to a size greater than the MSI spacial resolution. This makes a segmentation approach that allocates each pixel into a clearly separated class an appropriate route to obtaining a concise overview of the data.

The image segmentation, based purely on spectral differentiation that is preserved following basis approximation, clearly separated distinct regions of the tissue area. Hepatocytes are extracted from the surrounding tissue (orange) which consists mostly of fibrotic connective tissue, with the majority of hepatocytes being assigned to the same cluster (green). The major histological features that are commonly seen within NASH diseased tissue are visible in this segmentation and the normally smooth appearance of the liver has been severely deformed by the ballooning hepatocytes which are separated by regions of fibrotic connective tissue. This confirms that regions identified from the histological examination of the H&E stained tissue section have chemical changes as well as physical.

Liver hepatocytes would not be individually visible on this scale but are clearly identifiable following enlargement due to NASH. Histological examination and other visualisation approaches[67] suggested that some of the hepatocytes may be regenerating, specifically, within the large hepatocyte cluster indicated in the

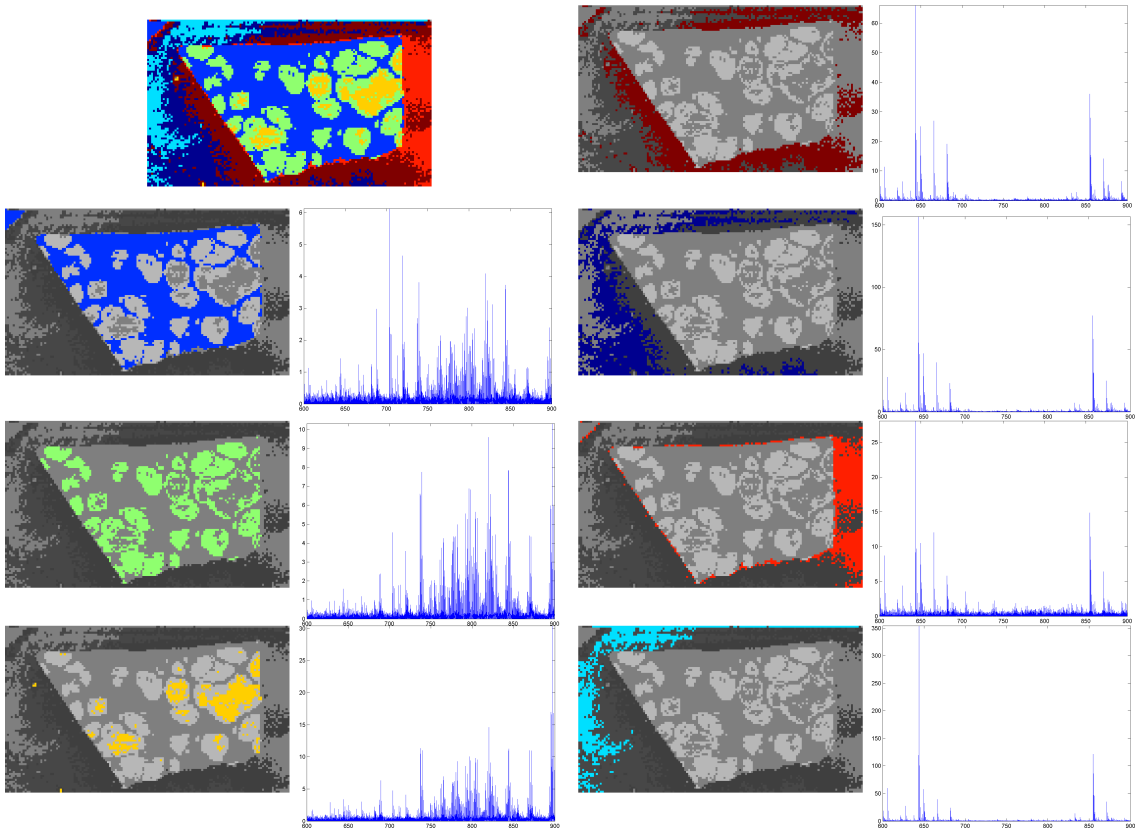

Figure 6.5: Compressed segmentation of a MSI dataset from a tissue section of diseased liver. Top left: segmentation map showing that all the tissue regions identified from histological examination were detected. Individually coloured images show the cluster localisation and the mean spectrum for each.

upper right of the image. This segmentation technique identified a subpopulation of hepatocytes (blue) which were thought to be regenerating nodules and selected the centre of these nodules as being spectrally distinct. All of these assignments are in agreement with the visualisation techniques used on these samples by *Fonville et al*[67]. This is notable as it takes at least two stains (typically H&E and Oil Red O[124]) to determine these structural and functional areas but MALDI MSI was able to discern them in a single experiment. The areas identified using segmentation can now form discrete regions of interest whose spectral differences can be examined. Underneath the segmentation map in Figure 6.5 are shown the average spectra for each cluster and differences in the magnitude of peaks are visible between the tissue regions. No discriminative peaks could be determined from comparison of the profiles which is not too surprising considering the cell types are very similar but the multivariate profiles reflect the differences in tissue environment.

6.3.3 Segmentation of Serial Sections

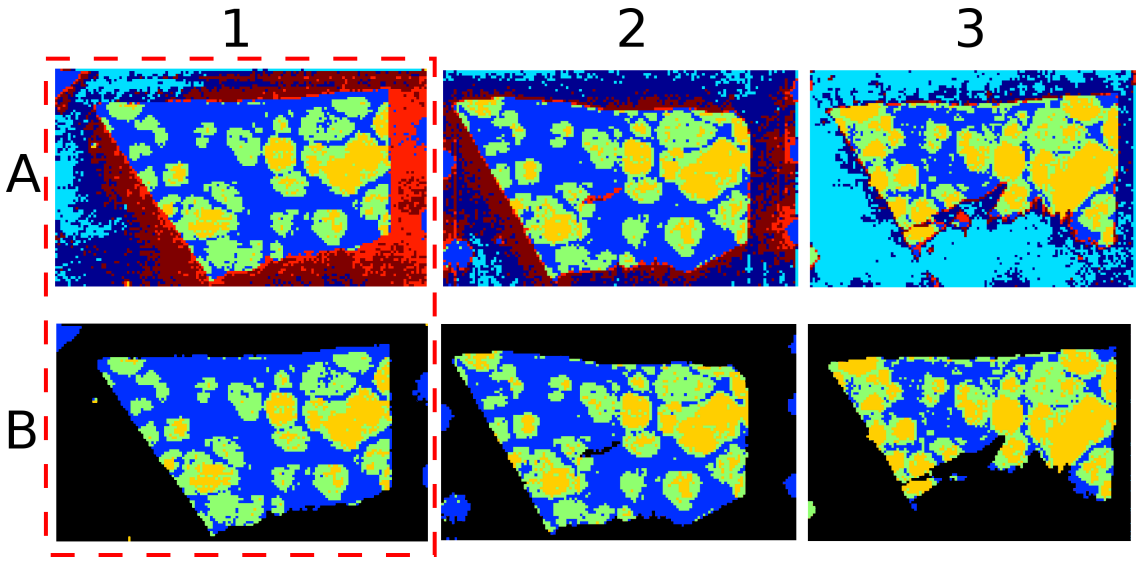

Figure 6.6: Classification of three MSI datasets collected from diseased liver sections using compressed segmentation (using spectral clustering) was performed. The clusters produced from dataset 1 were used to classify the two subsequent tissue sections. Row A shows the full segmentation, row B masks out the background for clarity of the on-tissue segmentation.

Two further MSI datasets were acquired from serial tissue sections of NASH diseased liver and the segmentation generated from the first section was used to identify the same histological regions in the subsequent two. A classification approach similar to that of the SOM was used. The average spectrum for each of the clusters was produced to provide 7 cluster centroids, back-projection with the randomised basis was used to recover the centroids in the m/z domain. Individual BASC models were constructed for each of the serial

tissue sections and the centroids were then projected into the basis of each, pixels in each dataset were then allocated to their most similar cluster (using the Euclidean distance). Another option would have been to perform a new spectral clustering after merging the datasets. Whilst these datasets could be combined using the basis merging algorithm, the resulting similarity matrix would become prohibitively large if all these pixels were included.

The resulting segmentation maps are shown in Figure 6.6. For clarity the clusters corresponding to background clusters (in the first image) have been set to black in Figure 6.6B. The dataset from the first tissue section was also re-classified using this centroid method to check the reliability of this classification approach. As the sections are serial the features visible should not change substantially between sections. Some stretching and deformation of the tissue sections occurred during collection and this is visible in the shape of the sections(the outline of which should be near identical) and the lower portion of the third section tore off during sectioning. The success of the approach can be seen from the segmentation maps, as spatially corresponding areas of tissue between the images are classified the same. As the sections were collected at 5 µm and the average mammalian cell is 10 µm then in many cases portions of the same cell were being analysed between the images so they should classify identically. The same background types appear in all images but with distributions that change, possibly reflecting differences in deposition of the MALDI matrix or contaminants on the plate. Some areas of contaminants (pen marks on the target plate) appear to be classified as tissue, however if the Euclidean distance between each pixel and its closest cluster centroid is plotted, see Figure 6.7 then it is clear that these are not very similar to the cluster they have been allocated to. An anomaly detection scheme could be incorporated to provide an automatic warning of inaccurate pixel segmentation or additional clusters incorporated within the clustering model.

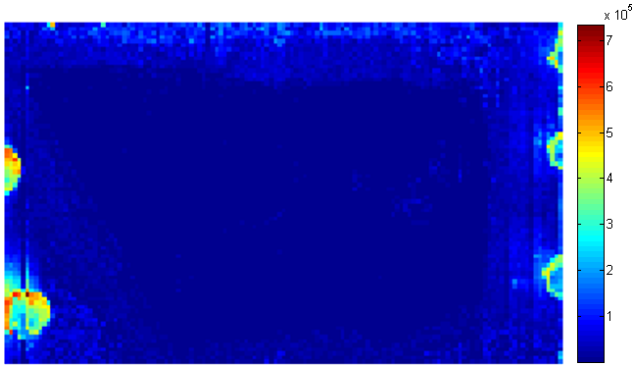

Figure 6.7: The distance to the allocated cluster threshold clearly identifies the mislabelled spatial anomalies present in the second and third datasets. Using such an anomaly detection metric could identify pixels mis-labelled during classification.

6.3.4 Conclusion

Mass spectrometry imaging has been shown to provide insights into the disease state of liver suffering from NASH and with the use of automated segmentation enlarged hepatocytes could be identified. The location of enlarged hepatocytes was confirmed by expert comparison to structural staining (H&E) but functional information about which nodules were potentially regenerating was also visible in the MSI data. Moreover, once the quality of segmentation had been confirmed by a domain expert the classification could be propagated to subsequent images. This highlights the potential for MALDI MSI as a routine tool for the analysis of tissue sections in a histopathological environment.

6.4 Conclusions

The interpretation of MSI data using basis approximation for compression and dimensionality reduction followed by segmentation using automated algorithms provided a clear overview of the tissue compartments within these datasets and fits functionally into existing workflows. In all cases speeds analysis compared to traditional workflows whilst preserving the full spectral resolution.

After defining spatial regions and extracting molecular profiles the final stage is in the elucidation and identification of molecules that are differentially detected between regions. The interpretation of the spatial maps still requires input from an appropriate expert but segmentation provides a way of presenting the results from mass spectrometry imaging in a format that can be readily understood by non-mass spectrometry experts. By using statistical methods to analyse the whole peaks whose variation corresponds to tissue compartments can be identified. This provides a simple approach for honing the list of peaks that require further analysis to only those with potential biological relevance. Analysis of the final images still requires the input from a biological domain expert but the hope is that these automated segmentation techniques will make the application of MSI something of a 'black box' for tissue analysis so that clinicians and biologists can apply it to addressing bio-medical questions.

6.5 Acknowledgements

Porcine ocular samples were provided by Roger Brown Ltd and the human liver samples were provided by Dr. Patricia Lalor (University of Birmingham). MALDI MSI of porcine ocular tissue and the human liver samples was performed within the research group of Dr. Josephine Bunch (University of Birmingham).

Chapter 7

Conclusions and Future Work

Matrix Assisted Laser Desorption Ionisation (MALDI) Mass Spectrometry (MS) imaging has become established as an important tool for the analysis of molecules directly from biological samples. Development of the underlying technology means Mass Spectrometry Imaging (MSI) in now routinely used in research and commercial environments and consequently more data than ever before is being produced. Datasets are continuing to increase in size either because the number of spectra are increasing through higher spatial resolution[236] or 3-dimensional datasets are being collected[161]. The methods developed within this thesis are therefore a timely addition to the MSI data analysis toolbox and provide a route to accelerating the processing, transfer and knowledge extraction from these highly complex datasets.

This thesis presented Basis Approximation for Spectral Compression (BASC) as a stand-alone step for dimensionality reduction and data compression following acquisition. The next stage is to consider where the most appropriate point for incorporating BASC into the MSI data collection workflow. As spectra are collected sequentially it would seem natural to try and incorporate online data processing in parallel with data acquisition as part of a standardised pipeline from sample preparation through acquisition and analysis. Online data processing is an important aspect of all the work here as it allows scaling in both number of pixels considered and the number of mass to charge ratio (m/z) values considered. As the sampling for the basis can be generated in parallel with data collection and multiple compressed portions of data can be combined post-compression this technique has potential for on-the-fly data compression and transmission. This may be particularly useful for spectrometers that are bandwidth limited in their transmission such as remote or field based devices. As miniaturisation of mass spectrometers continues and the spectral quality is improved it is natural to expect that a greater volume of data will be produced. These spectrometers

may have requirements of low-power or basic computational capacities having efficient methods that can be deployed locally could be essential.

The image compression and visualisation pipelines developed for MALDI-Time of Flight (TOF) MS have already been demonstrated on other hyper-spectral imaging modalities[165]. Different modalities collect complementary information about samples, for example MSI can provide elemental composition and Raman microscopy a quantitative measure of protein and lipid content, so it seems natural to take advantage of these information sources. Having methods that are readily translatable could allow laboratories to combine the information collected from these complementary modalities to advance the understanding of tissue processes and disease. As it appears to provide the most substantial savings when the spectral dimension greatly exceeds the number of spatial samplings it is natural to consider high mass-resolution imaging using Fourier Transform based detectors as another MSI field that could benefit from these methods as no other compression or factorisation technique has been demonstrated that preserves the full spectral resolution.

Robust assessment of the capabilities of automated data processing still requires further work for the validation of results. The simulation of datasets presented here has potential for further development, particularly if it is extended to high-resolution instrument types which can distinguish fine structure in isotope peaks and integrating models of ion formation or sample treatments like protein digests will be essential. The model was validated on an image with a limited (lipid) mass range. It would be a natural progression to validate the model across a larger mass range where different species are present (such as a linear TOF analyser). It may be preferable to characterise an instrument using a more controlled sample so that the spatial character of any noise can be considered. The development of analytical standards for MSI will greatly help in producing a thorough statistical model of the whole MALDI experiment[83]. Standardised samples and analysis pipelines should be developed based on statistically rigorous optimisation of experimental and analysis parameters so that they can be deployed with confidence.

Greater data sharing is required for the validation of image processing techniques between laboratories as it is currently almost impossible to investigate the reproducibility of methods presented in the literature. This thesis has suggested routes to overcoming this issue, either through compression using a BASC model or by sharing simulation parameters for synthetic datasets, but a community effort to standardise an approach is required before this will become mainstream. Projects such as openMSI[182] have taken steps towards putting imaging datasets into cloud storage and provided tools for visualisation of ion images. The extension of this platform to allow further data processing to be defined would be an excellent way to provide data in support of publications and shift the computational burden from local machines to large clusters. Having

a centralised community database for MSI would then permit the comparison of spectral profiles between different ionisation sources and instruments configurations and could lead to a substantial knowledge-base of tissue spectral profiles. Machine learning for tumour classification from a database of MSI spectral profiles has been developed for Desorption Electo-Spray Ionisation (DESI) [13, 219] and as MALDI is much more widespread it would be very interesting to see if this approach could be applied here too.

The technological and computational methods development for MSI should be validated to the point that they can be combined in a way that makes MSI something of a 'black box' for non-experts in MS. This would make the technique accessible to researchers seeking to answer specific biological, medical or pharmaceutical questions and allow the technology to become the powerful tool it has the potential to be.

Bibliography

[1] D. Achlioptas, *Database-friendly random projections: Johnson-lindenstrauss with binary coins*, Journal of computer and System Sciences, 66 (2003), pp. 671–687.

[2] T. Alexandrov, *Maldi imaging mass spectrometry: statistical data analysis and current computational challenges*, BMC bioinformatics, 13 (2012), pp. 1–13.

[3] T. Alexandrov and A. Bartels, *Testing for presence of known and unknown molecules in imaging mass spectrometry*, Bioinformatics, 29 (2013), pp. 2335–2342.

[4] T. Alexandrov, M. Becker, S. Deininger, G. Ernst, L. Wehder, M. Grasmair, F. von Eggeling, H. Thiele, and P. Maass, *Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering*, Journal of proteome research, 9 (2010), pp. 6535–6546.

[5] T. Alexandrov and J. H. Kobarg, *Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering*, Bioinformatics, 27 (2011), pp. i230–i238.

[6] T. Alexandrov and P. Lasch, *Segmentation of confocal raman microspectroscopic imaging data using edge-preserving denoising and clustering*, Analytical Chemistry, (2013).

[7] C. T. J. Alkemade, W. Snelleman, G. Boutilier, B. Pollard, J. Winefordner, T. Chester, and N. Omenetto, *A review and tutorial discussion of noise and signal-to-noise ratios in analytical spectrometryi. fundamental principles of signal-to-noise ratios*, Spectrochimica Acta Part B: Atomic Spectroscopy, 33 (1978), pp. 383–399.

[8] D. M. Anderson, D. Mills, J. Spraggins, W. S. Lambert, D. J. Calkins, and K. L. Schey, *High-resolution matrix-assisted laser desorption ionization–imaging mass spectrometry of lipids in rodent optic nerve tissue*, Molecular vision, 19 (2013), p. 581.

[9] P. Angulo, J. Keach, K. Batts, and K. Lindor, *Independent predictors of liver fibrosis in patients with nonalcoholic steatohepatitis*, Hepatology, 30 (1999), pp. 1356–1362.

[10] D. Arthur and S. Vassilvitskii, *k-means++: The advantages of careful seeding*, in Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[11] M. Babizhayev and A. Deyew, *Lens opacity induced by lipid peroxidation products as a model of cataract associated with retinal disease*, Biochimica et Biophysica Acta (BBA)-Lipids and Lipid Metabolism, 1004 (1989), pp. 124–133.

[12] J. Balog, L. Sasi-Szabó, J. Kinross, M. R. Lewis, L. J. Muirhead, K. Veselkov, R. Mirnezami, B. Dezső, L. Damjanovich, A. Darzi, et al., *Intraoperative tissue identification using rapid evaporative ionization mass spectrometry*, Science translational medicine, 5 (2013), pp. 194ra93–194ra93.

[13] J. Balog, T. Szaniszlo, K.-C. Schaefer, J. Denes, A. Lopata, L. Godorhazy, D. Szalay, L. Balogh, L. Sasi-Szabo, M. Toth, et al., *Identification of biological tissues by rapid evaporative ionization mass spectrometry*, Analytical chemistry, 82 (2010), pp. 7343–7350.

[14] D. Baluya, T. Garrett, and R. Yost, *Automated maldi matrix deposition method with inkjet printing for imaging mass spectrometry*, Analytical chemistry, 79 (2007), pp. 6862–6867.

[15] A. Bartels, P. Dülk, D. Trede, T. Alexandrov, and P. Maass, *Compressed sensing in imaging mass spectrometry*, Inverse Problems, 29 (2013), p. 125015.

[16] H. Basevi, K. Tichauer, F. Leblond, H. Dehghani, J. Guggenheim, R. Holt, and I. Styles, *Compressive sensing based reconstruction in bioluminescence tomography improves image resolution and robustness to noise*, Biomedical Optics Express, 3 (2012), pp. 2131–2141.

[17] E. Bingham and H. Mannila, *Random projection in dimensionality reduction: applications to image and text data*, in Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2001, pp. 245–250.

[18] E. Q. Blatherwick, C. I. Svensson, B. G. Frenguelli, and J. H. Scrivens, *Localisation of adenine nucleotides in heat-stabilised mouse brains using ion mobility enabled maldi imaging*, International Journal of Mass Spectrometry, (2013).

[19] D. Bonnel, R. Longuespee, J. Franck, M. Roudbaraki, P. Gosset, R. Day, M. Salzet, and I. Fournier, *Multivariate analyses for biomarkers hunting and validation through on-tissue bottom-up or in-source decay in maldi-msi: application to prostate cancer*, Analytical and Bioanalytical Chemistry, (2011), pp. 1–17.

[20] W. Bouschen and B. Spengler, *Artifacts of maldi sample preparation investigated by high-resolution scanning microprobe matrix-assisted laser desorption/ionization (smaldi) imaging mass spectrometry*, International Journal of Mass Spectrometry, 266 (2007), pp. 129–137.

[21] C. Bouveyron, S. Girard, and C. Schmid, *High-dimensional data clustering*, Computational Statistics & Data Analysis, 52 (2007), pp. 502–519.

[22] S. G. Boxer, M. L. Kraft, and P. K. Weber, *Advances in imaging secondary ion mass spectrometry for biological samples*, Annual review of biophysics, 38 (2009), pp. 53–74.

[23] R. N. Bracewell and R. Bracewell, *The Fourier transform and its applications*, vol. 31999, McGraw-Hill New York, 1986.

[24] J. Broersen, *Visualization in Large Scale Imaging Mass Spectrometry*, 2009.

[25] J. Bruand, T. Alexandrov, S. Sistla, M. Wisztorski, C. Meriaux, M. Becker, M. Salzet, I. Fournier, E. Macagno, and V. Bafna, *Amass: Algorithm for msi analysis by semi-supervised segmentation*, Journal of proteome research, 10 (2011), pp. 4734–4743.

[26] B. H. Bryant and K. Boekelheide, *Time-dependent changes in post-mortem testis histopathology in the rat*, Toxicologic pathology, 35 (2007), pp. 665–671.

[27] M. Bucknall, K. Y. Fung, and M. W. Duncan, *Practical quantitative biomedical applications of maldi-tof mass spectrometry*, Journal of the American Society for Mass Spectrometry, 13 (2002), pp. 1015–1027.

[28] J. Bunch, M. Clench, and D. Richards, *Determination of pharmaceutical compounds in skin by imaging matrix-assisted laser desorption/ionisation mass spectrometry*, Rapid communications in mass spectrometry, 18 (2004), pp. 3051–3060.

[29] C. Burrus and T. W. Parks, *DFT/FFT and Convolution Algorithms: theory and Implementation*, John Wiley & Sons, Inc., 1991.

[30] R. L. Caldwell and R. M. Caprioli, *Tissue profiling by mass spectrometry a review of methodology and applications*, Molecular & Cellular Proteomics, 4 (2005), pp. 394–401.

[31] R. M. Caprioli, T. B. Farmer, and J. Gile, *Molecular imaging of biological samples: localization of peptides and proteins using maldi-tof ms*, Analytical Chemistry, 69 (1997), pp. 4751–4760.

[32] C. Carter, C. McLeod, and J. Bunch, *Imaging of phospholipids in formalin fixed rat brain sections by matrix assisted laser desorption/ionization mass spectrometry*, Journal of the American Society for Mass Spectrometry, (2011), pp. 1–8.

[33] P. Chaurand, *Imaging mass spectrometry of thin tissue sections: A decade of collective efforts*, Journal of Proteomics, (2012).

[34] P. Chaurand, D. Cornett, P. Angel, and R. Caprioli, *From whole-body sections down to cellular level, multiscale imaging of phospholipids by maldi mass spectrometry*, Molecular & Cellular Proteomics, 10 (2011).

[35] P. Chaurand, J. L. Norris, D. S. Cornett, J. A. Mobley, and R. M. Caprioli, *New developments in profiling and imaging of proteins from tissue sections by maldi mass spectrometry*, Journal of proteome research, 5 (2006), pp. 2889–2900.

[36] P. Chaurand, S. Schwartz, and R. Caprioli, *Assessing protein patterns in disease using imaging mass spectrometry*, Journal of proteome research, 3 (2004), pp. 245–252.

[37] W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, and E. Y. Chang, *Parallel spectral clustering in distributed systems*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 33 (2011), pp. 568–586.

[38] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang, *The properties of high-dimensional data spaces: implications for exploring gene and protein expression data*, Nature Reviews Cancer, 8 (2008), pp. 37–49.

[39] L. H. Cohen and A. I. Gusev, *Small molecule analysis by maldi mass spectrometry*, Analytical and bioanalytical chemistry, 373 (2002), pp. 571–586.

[40] P. Comon, *Independent component analysis*, Higher-Order Statistics, (1992), pp. 29–38.

[41] K. Coombes, K. Baggerly, and J. Morris, *Pre-processing mass spectrometry data*, Fundamentals of Data Mining in Genomics and Proteomics, (2007), pp. 79–102.

[42] K. Coombes, S. Tsavachidis, J. Morris, K. Baggerly, M. Hung, and H. 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 (2005), pp. 4107–4117.

[43] K. R. Coombes, J. M. Koomen, K. A. Baggerly, J. S. Morris, and R. Kobayashi, *Understanding the characteristics of mass spectrometry data through the use of simulation*, Cancer Informatics, 1 (2005), p. 41.

[44] D. H. Cormack, *Essential histology*, Lippincott Williams & Wilkins, 2001.

[45] D. S. Cornett, J. A. Mobley, E. C. Dias, M. Andersson, C. L. Arteaga, M. E. Sanders, and R. M. Caprioli, *A novel histology-directed strategy for maldi-ms tissue profiling that improves throughput and cellular specificity in human breast cancer*, Molecular & Cellular Proteomics, 5 (2006), pp. 1975–1983.

[46] A. Crecelius, D. Cornett, R. Caprioli, B. Williams, B. Dawant, and B. Bodenheimer, *Three-dimensional visualization of protein expression in mouse brain structures using imaging mass spectrometry*, Journal of the American Society for Mass Spectrometry, 16 (2005), pp. 1093–1099.

[47] A. Cruz-Marcelo, R. Guerra, M. Vannucci, Y. Li, C. Lau, and T. Man, *Comparison of algorithms for pre-processing of seldi-tof mass spectrometry data*, Bioinformatics, 24 (2008), pp. 2129–2136.

[48] L. Da Vinci et al., *Leonardo on art and the artist*, DoverPublications. com, 1961.

[49] S. Dasgupta and A. Gupta, *An elementary proof of the johnson-lindenstrauss lemma*, tech. rep., Citeseer, 1999.

[50] A. M. De Marzo, H. H. Fedor, W. R. Gage, and M. A. Rubin, *Inadequate formalin fixation decreases reliability of p27¡ sup¿ kip1¡/sup¿ immunohistochemical staining: Probing optimal fixation time using high-density tissue microarrays*, Human pathology, 33 (2002), pp. 756–760.

[51] J. Deeley, J. Hankin, M. Friedrich, R. Murphy, R. Truscott, T. Mitchell, and S. Blanksby, *Sphingolipid distribution changes with age in the human lens*, Journal of lipid research, 51 (2010), pp. 2753–2760.

[52] S. Deininger, M. Becker, and D. Suckau, *Tutorial: multivariate statistical treatment of imaging data for clinical biomarker discovery*, Methods in Molecular Biology, 656 (2010), pp. 385–403.

[53] S. Deininger, D. Cornett, R. Paape, M. Becker, C. Pineau, S. Rauser, A. Walch, and E. Wolski, *Normalization in maldi-tof imaging datasets of proteins: practical considerations*, Analytical and Bioanalytical Chemistry, (2011), pp. 1–15.

[54] S. Deininger, M. Ebert, A. Fütterer, M. Gerhard, and C. Röcken, *Maldi imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers*, Journal of proteome research, 7 (2008), pp. 5230–5236.

[55] C. Ding, T. Li, W. Peng, and H. Park, *Orthogonal nonnegative matrix t-factorizations for clustering*, in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2006, pp. 126–135.

[56] M. Djidja, E. Claude, M. Snel, S. Francese, P. Scriven, V. Carolan, and M. Clench, *Novel molecular tumour classification using maldi–mass spectrometry imaging of tissue micro-array*, Analytical and bioanalytical chemistry, 397 (2010), pp. 587–601.

[57] K. Dreisewerd et al., *The desorption process in maldi*, Chemical Reviews-Columbus, 103 (2003), pp. 395–426.

[58] D. M. Drexler, T. J. Garrett, J. L. Cantone, R. W. Diters, J. G. Mitroka, M. C. Prieto Conaway, S. P. Adams, R. A. Yost, and M. Sanders, *Utility of imaging mass spectrometry (ims) by matrix-assisted laser desorption ionization (maldi) on an ion trap mass spectrometer in the analysis of drugs and metabolites in biological tissues*, Journal of pharmacological and toxicological methods, 55 (2007), pp. 279–288.

[59] P. Du, W. Kibbe, and S. Lin, *Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching*, Bioinformatics, 22 (2006), pp. 2059–2065.

[60] P. Du, G. Stolovitzky, P. Horvatovich, R. Bischoff, J. Lim, and F. Suits, *A noise model for mass spectrometry based proteomics*, Bioinformatics, 24 (2008), pp. 1070–1077.

[61] Q. Du and J. Fowler, *Hyperspectral image compression using jpeg2000 and principal component analysis*, Geoscience and Remote Sensing Letters, IEEE, 4 (2007), pp. 201–205.

[62] R. Durrant and A. Kabán, *Compressed fisher linear discriminant analysis: Classification of randomly projected data*, in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2010, pp. 1119–1128.

[63] L. Fabrigar, D. Wegener, R. MacCallum, and E. Strahan, *Evaluating the use of exploratory factor analysis in psychological research.*, Psychological methods, 4 (1999), p. 272.

[64] G. Farrell and C. Larter, *Nonalcoholic fatty liver disease: from steatosis to cirrhosis*, Hepatology, 43 (2006), pp. S99–S112.

[65] X. Fern and C. Brodley, *Random projection for high dimensional data clustering: A cluster ensemble approach*, in MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-, vol. 20, 2003, p. 186.

[66] J. Fonville, C. Carter, O. Cloarec, J. Nicholson, J. Lindon, J. Bunch, and E. Holmes, *Robust data processing and normalization strategy for maldi mass spectrometric imaging*, Analytical Chemistry, (2011).

[67] J. M. Fonville, C. L. Carter, L. Pizarro, R. T. Steven, A. D. Palmer, R. L. Griffiths, P. F. Lalor, J. C. Lindon, J. K. Nicholson, E. Holmes, et al., *Hyperspectral visualization of mass spectrometry imaging data*, Analytical chemistry, 85 (2013), pp. 1415–1423.

[68] J. Fowler, Q. Du, W. Zhu, and N. Younan, *Classification performance of random-projection-based dimensionality reduction of hyperspectral imagery*, in Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009, vol. 5, IEEE, 2009, pp. V–76.

[69] J. E. Fowler, *Compressive-projection principal component analysis*, Image Processing, IEEE Transactions on, 18 (2009), pp. 2230–2242.

[70] J. H. Friedman, *On bias, variance, 0/1loss, and the curse-of-dimensionality*, Data mining and knowledge discovery, 1 (1997), pp. 55–77.

[71] L. Galli and S. Salzo, *Lossless hyperspectral compression using klt*, in Geoscience and Remote Sensing Symposium, 2004. IGARSS'04. Proceedings. 2004 IEEE International, vol. 1, IEEE, 2004.

[72] E. Gaussier and C. Goutte, *Relation between plsa and nmf and implications*, in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2005, pp. 601–602.

[73] F. Gerber, F. Marty, G. B. Eijkel, K. Basler, E. Brunner, R. Furrer, and R. M. Heeren, *Multiorder correction algorithms to remove image distortions from mass spectrometry imaging data sets*, Analytical chemistry, 85 (2013), pp. 10249–10254.

[74] D. Gode and D. A. Volmer, *Lipid imaging by mass spectrometry–a review*, Analyst, 138 (2013), pp. 1289–1315.

[75] N. Goel, G. Bebis, and A. Nefian, *Face recognition experiments with random projection*, in Defense and Security, International Society for Optics and Photonics, 2005, pp. 426–437.

[76] R. J. Goodwin, *Sample preparation for mass spectrometry imaging: Small mistakes can lead to big consequences*, Journal of proteomics, 75 (2012), pp. 4893–4911.

[77] R. J. Goodwin, A. M. Lang, H. Allingham, M. Borén, and A. R. Pitt, *Stopping the clock on proteomic degradation by heat treatment at the point of tissue excision*, Proteomics, 10 (2010), pp. 1751–1761.

[78] L. Gorlitz, B. Menze, B. Kelm, and F. Hamprecht, *Processing spectral data*, (2009).

[79] A. Gowen, F. Marini, C. Esquerre, C. O´Donnell, G. Downey, and J. Burger, *Time series hyperspectral chemical imaging data: Challenges, solutions and applications*, Analytica chimica acta, 705 (2011), pp. 272–282.

[80] D. L. Gresh, *Self-corrected perceptual colormaps*, 2010.

[81] R. Griffiths and J. Bunch, *A survey of useful salt additives in matrix-assisted laser desorption/ionization mass spectrometry and tandem mass spectrometry of lipids: introducing nitrates for improved analysis*, Rapid Communications in Mass Spectrometry, 26 (2012), pp. 1557–1566.

[82] R. L. Griffiths, J. Sarsby, E. J. Guggenheim, A. M. Race, R. T. Steven, J. Fear, P. F. Lalor, and J. Bunch, *Formal lithium fixation improves direct analysis of lipids in tissue by mass spectrometry*, Analytical Chemistry, (2013).

[83] M. R. Groseclose and S. Castellino, *A mimetic tissue model for the quantification of drug distributions by maldi imaging mass spectrometry*, Analytical chemistry, 85 (2013), pp. 10099–10106.

[84] M. Guilhaus, D. Selby, and V. Mlynski, *Orthogonal acceleration time-of-flight mass spectrometry*, Mass spectrometry reviews, 19 (2000), pp. 65–107.

[85] N. Halko, P. Martinsson, and J. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM review, 53 (2011), pp. 217–288.

[86] G. Hamerly and C. Elkan, *Alternatives to the k-means algorithm that find better clusterings*, in Conference on Information and Knowledge Management: Proceedings of the eleventh international conference on Information and knowledge management, vol. 4, 2002, pp. 600–607.

[87] M. Hanselmann, M. Kirchner, B. Renard, E. Amstalden, K. Glunde, R. Heeren, and F. Hamprecht, *Concise representation of mass spectrometry images by probabilistic latent semantic analysis*, Analytical chemistry, 80 (2008), pp. 9649–9658.

[88] M. Hanselmann, U. Kothe, M. Kirchner, B. Renard, E. Amstalden, K. Glunde, R. Heeren, and F. Hamprecht, *Toward digital staining using imaging mass spectrometry and random forests*, Journal of proteome research, 8 (2009), pp. 3558–3567.

[89] M. Hanselmann, J. Roder, U. Kothe, B. Y. Renard, R. M. Heeren, and F. A. Hamprecht, *Active learning for convenient annotation and classification of secondary ion mass spectrometry images*, Analytical chemistry, 85 (2012), pp. 147–155.

[90] T. Hayasaka, N. Goto-Inoue, N. Zaima, K. Shrivas, Y. Kashiwagi, M. Yamamoto, M. Nakamoto, and M. Setou, *Imaging mass spectrometry with silver nanoparticles reveals the distribution of fatty acids in mouse retinal sections*, Journal of the American Society for Mass Spectrometry, 21 (2010), pp. 1446–1454.

[91] A. Henderson, J. Fletcher, and J. Vickerman, *A comparison of pca and maf for tof-sims image interpretation*, Surface and Interface Analysis, 41 (2009), pp. 666–674.

[92] K. R. Heys, M. G. Friedrich, and R. J. Truscott, *Free and bound water in normal and cataractous human lenses*, Investigative ophthalmology & visual science, 49 (2008), pp. 1991–1997.

[93] E. Hoffmann, *Mass spectrometry - Principles and Applications (3rd Edition)*, Wiley Online Library, 2007.

[94] S. Holt and R. M. Hicks, *Studies on formalin fixation for electron microscopy and cytochemical staining purposes*, The Journal of biophysical and biochemical cytology, 11 (1961), pp. 31–45.

[95] P. O. Hoyer, *Non-negative matrix factorization with sparseness constraints*, The Journal of Machine Learning Research, 5 (2004), pp. 1457–1469.

[96] S.-Y. Hsieh, C.-L. Tseng, Y.-S. Lee, A.-J. Kuo, C.-F. Sun, Y.-H. Lin, and J.-K. Chen, *Highly efficient classification and identification of human pathogenic bacteria by maldi-tof ms*, Molecular & cellular proteomics, 7 (2008), pp. 448–456.

[97] N. Hurley and S. Rickard, *Comparing measures of sparsity*, Information Theory, IEEE Transactions on, 55 (2009), pp. 4723–4741.

[98] A. Hyvarinen, *Fast ica for noisy data using gaussian moments*, in Circuits and Systems, 1999. ISCAS'99. Proceedings of the 1999 IEEE International Symposium on, vol. 5, IEEE, 1999, pp. 57–61.

[99] P. Indyk and R. Motwani, *Approximate nearest neighbors: towards removing the curse of dimensionality*, in Proceedings of the thirtieth annual ACM symposium on Theory of computing, ACM, 1998, pp. 604–613.

[100] A. Ipsen and T. M. Ebbels, *Prospects for a statistical theory of lc/tofms data*, Journal of the American Society for Mass Spectrometry, 23 (2012), pp. 779–791.

[101] A. K. Jain, *Data clustering: 50 years beyond k-means*, Pattern Recognition Letters, 31 (2010), pp. 651–666.

[102] D. Jameson and L. M. Hurvich, *Theory of brightness and color contrast in human vision*, Vision Research, 4 (1964), pp. 135–154.

[103] J. C. Jennette.

[104] W. Johnson and J. Lindenstrauss, *Extensions of lipschitz mappings into a hilbert space*, Contemporary mathematics, 26 (1984), pp. 1–1.

[105] I. Jolliffe and MyiLibrary, *Principal component analysis*, vol. 2, Wiley Online Library, 2002.

[106] A. W. Jones and P. Holmgren, *Uncertainty in estimating blood ethanol concentrations by analysis of vitreous humour*, Journal of clinical pathology, 54 (2001), pp. 699–702.

[107] E. Jones, S. Deininger, P. Hogendoorn, A. Deelder, and L. McDonnell, *Imaging mass spectrometry statistical analysis*, Journal of Proteomics, (2012).

[108] E. Jones, A. van Remoortere, R. van Zeijl, P. Hogendoorn, J. Bovée, A. Deelder, and L. McDonnell, *Multiple statistical analysis techniques corroborate intratumor heterogeneity in imaging mass spectrometry datasets of myxofibrosarcoma*, PloS one, 6 (2011), p. e24913.

[109] E. A. Jones, N. Schmitz, C. Waaijer, C. Frese, A. van Remoortere, R. van Zeijl, A. Heck, P. C. Hogendoorn, A. M. Deelder, M. Altelaar, et al., *Imaging mass spectrometry based molecular histology differentiates microscopically identical and heterogeneous tumors*, Journal of Proteome Research, (2013).

[110] J. H. Jungmann, L. MacAleese, J. Visser, M. J. Vrakking, and R. M. Heeren, *High dynamic range bio-molecular ion microscopy with the timepix detector*, Analytical chemistry, 83 (2011), pp. 7888–7894.

[111] J. Jurchen, S. Rubakhin, and J. Sweedler, *Maldi-ms imaging of features smaller than the size of the laser beam*, Journal of the American Society for Mass Spectrometry, 16 (2005), pp. 1654–1659.

[112] T. Kalvas, O. Tarvainen, T. Ropponen, O. Steczkiewicz, J. Arje, and H. Clark, *Ibsimu: A three-dimensional simulation software for charged particle optics*, Review of Scientific Instruments, 81 (2010), pp. 02B703–02B703.

[113] H. Kang, S. Lee, Y. Park, Y. Jeon, J. Lee, S.-Y. Jung, I. Park, S. Jang, H. Park, C. Yoo, et al., *Protein and lipid maldi profiles classify breast cancers according to the intrinsic subtype*, BMC cancer, 11 (2011), p. 465.

[114] M. Karas and F. Hillenkamp, *Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons*, Analytical chemistry, 60 (1988), pp. 2299–2301.

[115] S. Khatib-Shahidi, M. Andersson, J. Herman, T. Gillespie, and R. Caprioli, *Direct molecular analysis of whole-body animal tissue sections by imaging maldi mass spectrometry*, Analytical chemistry, 78 (2006), pp. 6448–6456.

[116] L. Klerk, A. Broersen, I. Fletcher, R. van Liere, and R. Heeren, *Extended data analysis strategies for high resolution imaging ms: New methods to deal with extremely large image hyperspectral datasets*, International journal of mass spectrometry, 260 (2007), pp. 222–236.

[117] R. Knochenmuss, *Ion formation mechanisms in uv-maldi*, Analyst, 131 (2006), pp. 966–986.

[118] R. Knochenmuss and R. Zenobi, *Maldi ionization: the role of in-plume processes*, Chemical reviews, 103 (2003), pp. 441–452.

[119] T. Kohonen, *The self-organizing map*, Proceedings of the IEEE, 78 (1990), pp. 1464–1480.

[120] H. Kojima, S. Sakurai, M. Uemura, T. Takekawa, H. Morimoto, Y. Tamagawa, and H. Fukui, *Difference and similarity between non-alcoholic steatohepatitis and alcoholic liver disease*, Alcoholism: Clinical and Experimental Research, 29 (2005), pp. 259S–263S.

[121] H. Kriegel, P. Kröger, and A. Zimek, *Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering*, ACM Transactions on Knowledge Discovery from Data (TKDD), 3 (2009), p. 1.

[122] A. N. Krutchinsky and B. T. Chait, *On the nature of the chemical noise in maldi mass spectra*, Journal of the American Society for Mass Spectrometry, 13 (2002), pp. 129–134.

[123] M. Lagarrigue, M. Becker, R. Lavigne, S. Deininger, A. Walch, F. Aubry, D. Suckau, and C. Pineau, *Revisiting rat spermatogenesis with maldi imaging at 20-μm resolution*, Molecular & Cellular Proteomics, 10 (2011).

[124] P. Lalor, J. Faint, Y. Aarbodem, S. Hubscher, D. Adams, et al., *The role of cytokines and chemokines in the development of steatohepatitis*, in Seminars in Liver Diseases, vol. 27, New York: Thieme-Stratton, c1981-, 2007, pp. 173–193.

[125] I. Lanekoff, M. Thomas, J. P. Carson, J. N. Smith, C. Timchalk, and J. Laskin, *Imaging nicotine in rat brain tissue by use of nanospray desorption electrospray ionization mass spectrometry*, Analytical chemistry, 85 (2013), pp. 882–889.

[126] E. S. Lein, M. J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A. F. Boe, M. S. Boguski, K. S. Brockway, E. J. Byrnes, et al., *Genome-wide atlas of gene expression in the adult mouse brain*, Nature, 445 (2006), pp. 168–176.

[127] R. Lemaire, A. Desmons, J. Tabet, R. Day, M. Salzet, and I. Fournier, *Direct analysis and maldi imaging of formalin-fixed, paraffin-embedded tissue sections*, Journal of proteome research, 6 (2007), pp. 1295–1305.

[128] K. C. Leptos, D. A. Sarracino, J. D. Jaffe, B. Krastins, and G. M. Church, *Mapquant: Open-source software for large-scale protein quantification*, Proteomics, 6 (2006), pp. 1770–1782.

[129] I. Levner, *Feature selection and nearest centroid classification for protein mass spectrometry*, BMC bioinformatics, 6 (2005), p. 68.

[130] L. Li, R. W. Garden, and J. V. Sweedler, *Single-cell maldi: a new tool for direct peptide profiling*, Trends in biotechnology, 18 (2000), pp. 151–160.

[131] Y. Li and A. Ngom, *The non-negative matrix factorization toolbox for biological data mining*, Source code for biology and medicine, 8 (2013), pp. 1–15.

[132] C. B. Lietz, E. Gemperline, and L. Li, *Qualitative and quantitative mass spectrometry imaging of drugs and metabolites*, Advanced Drug Delivery Reviews, (2013).

[133] J. Lin and D. Gunopulos, *Dimensionality reduction by random projection and latent semantic indexing*, in proceedings of the Text Mining Workshop, at the 3rd SIAM International Conference on Data Mining, 2003.

[134] J. Lindenstrauss and A. Pelczynski, *Absolutely summing operators in lp-spaces and their applications*, Studia Math, 29 (1968), p. 193.

[135] Q. Liu, A. Sung, M. Qiao, Z. Chen, J. Yang, M. Yang, X. Huang, and Y. Deng, *Comparison of feature selection and classification for maldi-ms data*, BMC genomics, 10 (2009), p. S3.

[136] G. R. Lloyd, C. Kendall, T. Cook, N. Shepherd, N. Stone, et al., *Histological imaging of a human colon polyp sample using raman spectroscopy and self organising maps*, Vibrational Spectroscopy, (2012).

[137] W. Lu and J. C. Rajapakse, *Eliminating indeterminacy in ica*, Neurocomputing, 50 (2003), pp. 271–290.

[138] V. Mainini, G. Bovo, C. Chinello, E. Gianazza, M. Grasso, G. Cattoretti, and F. Magni, *Detection of high molecular weight proteins by maldi imaging mass spectrometry*, Molecular BioSystems, (2013).

[139] A. Majumdar and R. K. Ward, *Robust classifiers for data reduced via random projections*, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 40 (2010), pp. 1359–1371.

[140] A. Malpica, M. T. Deavers, K. Lu, D. C. Bodurka, E. N. Atkinson, D. M. Gershenson, and E. G. Silva, *Grading ovarian serous carcinoma using a two-tier system*, The American journal of surgical pathology, 28 (2004), pp. 496–504.

[141] D. Mantini, F. Petrucci, P. Del Boccio, D. Pieragostino, M. Di Nicola, A. Lugaresi, G. Federici, P. Sacchetta, C. Di Ilio, and A. Urbani, *Independent component analysis for the extraction of reliable protein signal profiles from maldi-tof mass spectra*, Bioinformatics, 24 (2008), pp. 63–70.

[142] D. Mantini, F. Petrucci, D. Pieragostino, P. Del Boccio, M. Di Nicola, C. Di Ilio, G. Federici, P. Sacchetta, S. Comani, and A. Urbani, *Limpic: a computational method for the separation of protein maldi-tof-ms signals from noise*, BMC bioinformatics, 8 (2007), p. 101.

[143] D. Martin, C. Fowlkes, D. Tal, and J. Malik, *A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics*, in Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, vol. 2, IEEE, 2001, pp. 416–423.

[144] S. Mazaud-Guittot, E. Meugnier, S. Pesenti, X. Wu, H. Vidal, A. Gow, and B. Le Magueresse-Battistoni, *Claudin 11 deficiency in mice results in loss of the sertoli cell epithelial phenotype in the testis*, Biology of reproduction, 82 (2010), pp. 202–213.

[145] W. G. McCluggage, L. Hirschowitz, G. E. Wilson, E. Oliva, R. A. Soslow, and R. J. Zaino, *Significant variation in the assessment of cervical involvement in endometrial carcinoma: an interobserver variation study*, The American journal of surgical pathology, 35 (2011), pp. 289–294.

[146] G. McCombie, D. Staab, M. Stoeckli, and R. Knochenmuss, *Spatial and spectral correlations in maldi mass spectrometry images by clustering and multivariate analysis*, Analytical chemistry, 77 (2005), pp. 6118–6124.

[147] L. McDonnell, A. Van Remoortere, N. De Velde, R. Van Zeijl, and A. Deelder, *Imaging mass spectrometry data reduction: Automated feature identification and extraction*, Journal of the American Society for Mass Spectrometry, 21 (2010), pp. 1969–1978.

[148] L. McDonnell, A. van Remoortere, R. van Zeijl, H. Dalebout, M. Bladergroen, and A. Deelder, *Automated imaging ms: toward high throughput imaging mass spectrometry*, Journal of proteomics, 73 (2010), pp. 1279–1282.

[149] L. A. McDonnell, S. R. Piersma, A. Altelaar, T. H. Mize, S. L. Luxembourg, P. D. Verhaert, J. van Minnen, and R. Heeren, *Subcellular imaging mass spectrometry of brain tissue*, Journal of mass spectrometry, 40 (2005), pp. 160–168.

[150] S. Meding, U. Nitsche, B. Balluff, M. Elsner, S. Rauser, C. Schone, M. Nipp, M. Maak, M. Feith, M. P. Ebert, et al., *Tumor classification of six common cancer types based on proteomic profiling by maldi imaging*, Journal of proteome research, 11 (2012), pp. 1996–2003.

[151] T. Menzies and M. Shepperd, *Special issue on repeatable results in software engineering prediction*, Empirical Software Engineering, 17 (2012), pp. 1–17.

[152] A. L. Mescher, *Junqueira's basic histology: Text & atlas*, Mcgraw-hill medical New York, 2010.

[153] W. Meuleman, J. Engwegen, M. Gast, J. Beijnen, M. Reinders, and L. Wessels, *Comparison of normalisation methods for surface-enhanced laser desorption and ionisation (seldi) time-of-flight (tof) mass spectrometry data*, BMC bioinformatics, 9 (2008), p. 88.

[154] L. Miao and H. Qi, *Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization*, Geoscience and Remote Sensing, IEEE Transactions on, 45 (2007), pp. 765–777.

[155] M. Miettinen and J. Lasota, *Gastrointestinal stromal tumors: review on morphology, molecular pathology, prognosis, and differential diagnosis*, Archives of pathology & laboratory medicine, 130 (2006), pp. 1466–1478.

[156] A. Moore, *K-means and hierarchical clustering*, 2001.

[157] J. S. Morris, K. R. Coombes, J. Koomen, K. A. Baggerly, and R. Kobayashi, *Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum*, Bioinformatics, 21 (2005), pp. 1764–1775.

[158] R. Murphy, J. Hankin, and R. Barkley, *Imaging of lipid species by maldi mass spectrometry*, Journal of lipid research, 50 (2009), pp. S317–S322.

[159] E. Nordhoff, R. Cramer, M. Karas, F. Hillenkamp, F. Kirpekar, K. Kristiansen, and P. Roepstorff, *Ion stability of nucleic acids in infrared matrix-assisted laser desorption/ionization mass spectrometry*, Nucleic acids research, 21 (1993), pp. 3347–3357.

[160] H. Nygren, K. Börner, B. Hagenhoff, P. Malmberg, and J.-E. Månsson, *Localization of cholesterol, phosphocholine and galactosylceramide in rat cerebellar cortex with imaging tof-sims equipped with a bismuth cluster ion source*, Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids, 1737 (2005), pp. 102–110.

[161] J. Oetjen, M. Aichler, D. Trede, J. Strehlow, J. Berger, S. Heldmann, M. Becker, M. Gottschalk, J. H. Kobarg, S. Wirtz, et al., *Mri-compatible pipeline for three-dimensional maldi imaging mass spectrometry using paxgene fixation*, Journal of proteomics, (2013).

[162] E. Oja and M. Plumbley, *Blind separation of positive sources using nonnegative pca*, in 4th International Symposium on Independent Component Analysis and Blind Signal Separation, 2003.

[163] S. H. Ong, N. Yeo, K. Lee, Y. Venkatesh, and D. Cao, *Segmentation of color images using a two-stage self-organizing network*, Image and Vision Computing, 20 (2002), pp. 279–289.

[164] A. Palmer, R. Griffiths, I. Styles, E. Claridge, A. Calcagni, and J. Bunch, *Sucrose cryo-protection facilitates imaging of whole eye sections by maldi mass spectrometry*, Journal of Mass Spectrometry, 47 (2012), pp. 237–241.

[165] A. D. Palmer, A. Bannerman, L. Grover, and I. B. Styles, *Faster tissue interface analysis from raman microscopy images using compressed factorisation*, in European Conferences on Biomedical Optics, International Society for Optics and Photonics, 2013, pp. 87980H–87980H.

[166] A. D. Palmer, J. Bunch, and I. B. Styles, *Randomised approximation methods for the efficient compression and analysis of hyperspectral data*, Analytical chemistry, (2013).

[167] H. Pan, H. Zhang, D. Chang, H. Li, and H. Sui, *The change of oxidative stress products in diabetes mellitus and diabetic retinopathy*, British Journal of Ophthalmology, 92 (2008), pp. 548–551.

[168] J.-W. Park, H. Min, Y.-P. Kim, H. Kyong Shon, J. Kim, D. W. Moon, and T. G. Lee, *Multivariate analysis of tof-sims data for biological applications*, Surface and Interface Analysis, 41 (2009), pp. 694–703.

[169] K. Pearson, *Liii. on lines and planes of closest fit to systems of points in space*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2 (1901), pp. 559–572.

[170] W. Pitkeathly, S. Rezatofighi, J. Rappoport, and E. Claridge, *A framework for generating realistic synthetic sequences of dynamic confocal microscopy images*, BMVA, 2013.

[171] S. M. Puolitaival, K. E. Burnum, D. S. Cornett, and R. M. Caprioli, *Solvent-free matrix dry-coating for maldi imaging of phospholipids*, Journal of the American Society for Mass Spectrometry, 19 (2008), pp. 882–886.

[172] A. Race, R. Steven, A. Palmer, I. Styles, and J. Bunch, *Memory efficient principal component analysis for the dimensionality reduction of large mass spectrometry imaging datasets.*, Analytical chemistry, (2013).

[173] A. Race, I. Styles, and J. Bunch, *Inclusive sharing of mass spectrometry imaging data requires a converter for all*, Journal of Proteomics, (2012).

[174] A. M. Race, R. T. Steven, A. D. Palmer, I. B. Styles, and J. Bunch, *Memory efficient principal component analysis of large mass spectrometry imaging datasets*, OurCon Poster Presentation, 1 (2012).

[175] W. Reindl, B. P. Bowen, M. A. Balamotis, J. E. Green, and T. R. Northen, *Multivariate analysis of a 3d mass spectral image for examining tissue heterogeneity*, Integrative Biology, 3 (2011), pp. 460–467.

[176] B. Y. Renard, M. Kirchner, H. Steen, J. A. Steen, and F. A. Hamprecht, *Nitpick: peak identification for mass spectrometry data*, BMC bioinformatics, 9 (2008), p. 355.

[177] S. H. Rezatofighi, W. Pitkeathly, S. Gould, R. Hartley, K. Mele, W. Hughes, and J. Burchfield, *A framework for generating realistic synthetic sequences of total internal reflection fluorescence microscopy images*, in Proc. ISBI, 2013.

[178] M. Rojo, G. Bueno, and J. Slodkowska, *Review of imaging solutions for integrated quantitative immunohistochemistry in the pathology daily practice*, Folia Histochemica et Cytobiologica, 47 (2009), pp. 349–354.

[179] A. Römpp, S. Guenther, Z. Takats, and B. Spengler, *Mass spectrometry imaging with high resolution in mass and space (hr2 msi) for reliable investigation of drug compound distributions on the cellular level*, Analytical and bioanalytical chemistry, 401 (2011), pp. 65–73.

[180] M. Roy, H. Nakanishi, K. Takahashi, S. Nakanishi, S. Kajihara, T. Hayasaka, M. Setou, K. Ogawa, R. Taguchi, and T. Naito, *Salamander retina phospholipids and their localization by maldi imaging mass spectrometry at cellular size resolution*, Journal of Lipid Research, 52 (2011), pp. 463–470.

[181] S. Rubakhin and J. Sweedler, *Mass Spectrometry Imaging: Principles and Protocols*, Springer, 2010.

[182] O. R ubel, A. Greiner, S. Cholia, K. Louie, E. W. Bethel, T. R. Northen, and B. P. Bowen, *Openmsi: A high-performance web-based platform for mass spectrometry imaging*, Analytical chemistry, 85 (2013), pp. 10354–10361.

[183] A. C. Sauve and T. P. Speed, *Normalization, baseline correction and alignment of high-throughput mass spectrometry data*, Proceedings Gensips, (2004).

[184] A. Savitzky and M. Golay, *Smoothing and differentiation of data by simplified least squares procedures.*, Analytical chemistry, 36 (1964), pp. 1627–1639.

[185] M. Schuerenberg and S.-O. Deininger, *Matrix application with imageprep*, in Imaging Mass Spectrometry, Springer, 2010, pp. 87–91.

[186] A. SCIEX.

[187] M. Shariatgorji, A. Nilsson, R. J. Goodwin, P. Svenningsson, N. Schintu, Z. Banka, L. Kladni, T. Hasko, A. Szabo, and P. E. Andren, *Deuterated matrix-assisted laser desorption ionization matrix uncovers masked mass spectrometry imaging signals of small molecules*, Analytical chemistry, 84 (2012), pp. 7152–7157.

[188] A. Sherstyuk, *Kernel functions in convolution surfaces: a comparative analysis*, The Visual Computer, 15 (1999), pp. 171–182.

[189] S.-R. Shi, M. E. Key, and K. L. Kalra, *Antigen retrieval in formalin-fixed, paraffin-embedded tissues: an enhancement method for immunohistochemical staining based on microwave oven heating of tissue sections.*, Journal of Histochemistry & Cytochemistry, 39 (1991), pp. 741–748.

[190] Z. Shi, L. Liu, X. Zhai, and Z. Jiang, *Efficient sparse unmixing analysis for hyperspectral imagery based on random projection*, Neural Computing & Applications, (2012), pp. 1–13.

[191] S. Silva, B. Sousa Santos, and J. Madeira, *Using color in visualization: A survey*, Computers & Graphics, 35 (2011), pp. 320–333.

[192] P. W. Siy, R. A. Moffitt, R. M. Parry, Y. Chen, Y. Liu, M. C. Sullards, A. H. Merrill, and M. D. Wang, *Matrix factorization techniques for analysis of imaging mass spectrometry data*, in BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on, IEEE, 2008, pp. 1–6.

[193] I. Smirnov, X. Zhu, T. Taylor, Y. Huang, P. Ross, I. Papayanopoulos, S. Martin, and D. Pappin, *Suppression of α-cyano-4-hydroxycinnamic acid matrix clusters and reduction of chemical noise in maldi-tof mass spectrometry*, Analytical chemistry, 76 (2004), pp. 2958–2965.

[194] D. F. Smith, C. Schulz, M. Konijnenburg, M. Kilic, and R. M. Heeren, *Distributed computing strategies for processing of ft-icr ms imaging datasets for continuous mode data visualization*, Analytical and bioanalytical chemistry, (2014), pp. 1–7.

[195] R. Somorjai, B. Dolenko, and R. Baumgartner, *Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions*, Bioinformatics, 19 (2003), pp. 1484–1491.

[196] R. T. Steven and J. Bunch, *Repeat maldi ms imaging of a single tissue section using multiple matrices and tissue washes*, Analytical and bioanalytical chemistry, (2013), pp. 1–10.

[197] R. T. Steven, A. D. Palmer, and J. Bunch, *Fluorometric beam profiling of uv maldi lasers*, Journal of The American Society for Mass Spectrometry, pp. 1–7.

[198] R. T. Steven, A. M. Race, and J. Bunch, *para-nitroaniline is a promising matrix for maldi-ms imaging on intermediate pressure ms systems*, Journal of The American Society for Mass Spectrometry, (2013), pp. 1–4.

[199] M. Stoeckli, P. Chaurand, D. E. Hallahan, and R. M. Caprioli, *Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues*, Nature medicine, 7 (2001), pp. 493–496.

[200] M. Stoeckli, T. B. Farmer, and R. M. Caprioli, *Automated mass spectrometry imaging with a matrix-assisted laser desorption ionization time-of-flight instrument*, Journal of the American Society for Mass Spectrometry, 10 (1999), pp. 67–71.

[201] G. Stone, D. Clifford, J. Gustafsson, S. McColl, and P. Hoffmann, *Visualisation in imaging mass spectrometry using the minimum noise fraction transform*, BMC Research Notes, 5 (2012), p. 419.

[202] M. Strohalm, J. Strohalm, F. Kaftan, L. Krasny, M. Volny, P. Novak, K. Ulbrich, and V. Havlicek, *Poly [n-(2-hydroxypropyl) methacrylamide]-based tissue-embedding medium compatible with maldi mass spectrometry imaging experiments*, Analytical chemistry, 83 (2011), pp. 5458–5462.

[203] F. Suits, T. Fehniger, A. Vegvari, G. Marko-Varga, and P. Horvatovich, *Correlation queries for mass spectrometry imaging*, Analytical chemistry, (2013).

[204] V. Sulic, J. Perš, M. Kristan, and S. Kovacic, *Efficient dimensionality reduction using random projection*, in Computer Vision Winter Workshop, 2010.

[205] E. Sweeney, T. H. Ward, N. Gray, C. Womack, G. Jayson, A. Hughes, C. Dive, and R. Byers, *Quantitative multiplexed quantum dot immunohistochemistry*, Biochemical and biophysical research communications, 374 (2008), pp. 181–186.

[206] I. Taban, A. Altelaar, Y. van der Burgt, L. McDonnell, R. Heeren, J. Fuchser, and G. Baykut, *Imaging of peptides in the rat brain using maldi-fticr mass spectrometry*, Journal of the American Society for Mass Spectrometry, 18 (2007), pp. 145–151.

[207] K. Tanaka, H. Waki, Y. Ido, S. Akita, Y. Yoshida, T. Yoshida, and T. Matsuo.

[208] X. Tang, W. Pearlman, and J. Modestino, *Hyperspectral image compression using three-dimensional wavelet coding*, in IEEE Int. Conf. Image Process., 2004, pp. 1133 –1136.

[209] H. Thiele, S. Heldmann, D. Trede, J. Strehlow, S. Wirtz, W. Dreher, J. Berger, J. Oetjen, J. H. Kobarg, B. Fischer, et al., *2d and 3d maldi-imaging: conceptual strategies for visualization and data mining*, Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 1844 (2014), pp. 117–137.

[210] D. Trede, S. Schiffler, M. Becker, S. Wirtz, K. Steinhorst, J. Strehlow, M. Aichler, J. Kobarg, J. Oetjen, A. Dyatlov, et al., *Exploring three-dimensional matrix-assisted laser desorption/ionization imaging mass spectrometry data: Three-dimensional spatial segmentation of mouse kidney*, Analytical chemistry, 84 (2012), pp. 6079–6087.

[211] P. Trim, M. Djidja, S. Atkinson, K. Oakes, L. Cole, D. Anderson, P. Hart, S. Francese, and M. Clench, *Introduction of a 20 khz nd: Yvo4 laser into a hybrid quadrupole time-of-flight mass spectrometer for maldi-ms imaging*, Analytical and bioanalytical chemistry, 397 (2010), pp. 3409–3419.

[212] B. Tyler, *Interpretation of tof-sims images: multivariate and univariate approaches to image de-noising, image segmentation and compound identification*, Applied surface science, 203 (2003), pp. 825–831.

[213] B. Tyler, G. Rayal, and D. Castner, *Multivariate analysis strategies for processing tof-sims images of biomaterials*, Biomaterials, 28 (2007), pp. 2412–2423.

[214] B. J. Tyler and R. 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 (2013), pp. 475–478.

[215] R. Van de Plas, B. De Moor, and E. Waelkens, *Discrete wavelet transform-based multivariate exploration of tissue via imaging mass spectrometry*, in proceedings of the 2008 ACM symposium on applied computing, ACM, 2008, pp. 1307–1308.

[216] L. Van der Maaten and G. Hinton, *Visualizing data using t-sne*, Journal of Machine Learning Research, 9 (2008), p. 85.

[217] K. Varmuza, C. Engrand, P. Filzmoser, M. Hilchenbach, J. Kissel, H. Krüger, J. Silén, and M. Trieloff, *Random projection for dimensionality reductionapplied to time-of-flight secondary ion mass spectrometry data*, Analytica chimica acta, 705 (2011), pp. 48–55.

[218] K. Varmuza, P. Filzmoser, and B. Liebmann, *Random projection experiments with chemometric data*, Journal of Chemometrics, 24 (2010), pp. 209–217.

[219] K. A. Veselkov, R. Mirnezami, N. Strittmatter, R. D. Goldin, J. Kinross, A. V. Speller, T. Abramov, E. A. Jones, A. Darzi, E. Holmes, et al., *Chemo-informatic strategy for imaging mass spectrometry-based hyperspectral profiling of lipid signatures in colorectal cancer*, Proceedings of the National Academy of Sciences, 111 (2014), pp. 1216–1221.

[220] U. Von Luxburg, *A tutorial on spectral clustering*, Statistics and computing, 17 (2007), pp. 395–416.

[221] A. Walch, S. Rauser, S.-O. Deininger, and H. Höfler, *Maldi imaging mass spectrometry for direct tissue analysis: a new frontier for molecular histology*, Histochemistry and cell biology, 130 (2008), pp. 421–434.

[222] J. Watrous, P. J. Roach, B. S. Heath, T. Alexandrov, J. Laskin, and P. C. Dorrestein, *Metabolic profiling directly from the petri dish using nanodesi imaging mass spectrometry*, Analytical Chemistry, (2013).

[223] J. D. Watrous, T. Alexandrov, and P. C. Dorrestein, *The evolving field of imaging mass spectrometry and its impact on future biological research*, Journal of Mass Spectrometry, 46 (2011), pp. 209–222.

[224] B. Welford, *Note on a method for calculating corrected sums of squares and products*, Technometrics, 4 (1962), pp. 419–420.

[225] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, *Feature selection for svms*, in NIPS, vol. 12, 2000, pp. 668–674.

[226] B. Williams, S. Cornett, B. Dawant, A. Crecelius, B. Bodenheimer, and R. Caprioli, *An algorithm for baseline correction of maldi mass spectra*, in Proceedings of the 43rd annual Southeast regional conference, vol. 1, 2005, pp. 137–142.

[227] M. Wisztorski, J. Franck, M. Salzet, and I. Fournier, *Maldi direct analysis and imaging of frozen versus ffpe tissues: what strategy for which sample?*, in Mass Spectrometry Imaging, Springer, 2010, pp. 303–322.

[228] M. Wolkenstein, H. Hutter, C. Mittermayr, W. Schiesser, and M. Grasserbauer, *Classification of sims images using a kohonen network*, Analytical Chemistry, 69 (1997), pp. 777–782.

[229] M. Wolkenstein, T. Stubbings, and H. Hutter, *Robust automated three-dimensional segmentation of secondary ion mass spectrometry image sets*, Fresenius' journal of analytical chemistry, 365 (1999), pp. 63–69.

[230] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao, *Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data*, Bioinformatics, 19 (2003), pp. 1636–1643.

[231] W. Xu, X. Liu, and Y. Gong, *Document clustering based on non-negative matrix factorization*, in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM, 2003, pp. 267–273.

[232] W. Yan, J. A. Gardella Jr, and T. D. Wood, *Quantitative analysis of technical polymer mixtures by matrix assisted laser desorption/ionization time of flight mass spectrometry*, Journal of the American Society for Mass Spectrometry, 13 (2002), pp. 914–920.

[233] C. Yang, Z. He, and W. Yu, *Comparison of public peak detection algorithms for maldi mass spectrometry data analysis*, BMC bioinformatics, 10 (2009), p. 4.

[234] J. Yang, J. Wright, T. Huang, and Y. Ma, *Image super-resolution as sparse representation of raw image patches*, in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, 2008, pp. 1–8.

[235] C. Zavaleta, B. Smith, I. Walton, W. Doering, G. Davis, B. Shojaei, M. Natan, and S. Gambhir, *Multiplexed imaging of surface enhanced raman scattering nanotags in living mice using noninvasive raman spectroscopy*, Proceedings of the National Academy of Sciences, 106 (2009), pp. 13511–13516.

[236] A. Zavalin, J. Yang, and R. Caprioli, *Laser beam filtration for high spatial resolution maldi imaging mass spectrometry*, Journal of The American Society for Mass Spectrometry, 24 (2013), pp. 1153–1156.

[237] J. Zhang, J. Erway, X. Hu, Q. Zhang, and R. Plemmons, *Randomized svd methods in hyperspectral imaging*, Journal of Electrical and Computer Engineering, 2012 (2012).