Statistical Treatment of Proteomic Imaging Mass Spectrometry Data

by Lyron Juan Winderbaum

Primary Supervisor: Associate Professor Inge Koch

Co - Supervisor: Professor Peter Hoffmann

Thesis submitted for the degree of Doctor of Philosophy

June, 2016

DISCIPLINE OF STATISTICS
SCHOOL OF MATHEMATICAL SCIENCES

THE UNIVERSITY of ADELAIDE

Contents

Acronyms

m/z Mass-to-Charge Ratio.

APC Adelaide Proteomics Centre.

CCA Canonical Correlation Analysis.

CV Cross Validation.

Da Dalton.

DIPPS Difference in Proportions of Occurrence Statistic.

DWD Distance Weighted Discrimination.

FFPE Formalin Fixed and Paraffin Embedded.

GE Gel Electrophoresis.

H&E Hematoxylin and Eosin.

HDLSS High-Dimension Low Sample Size.

IHC Immunohistochemistry.

LC Liquid Chromatography.

LDA Linear Discriminant Analysis.

LNM Lymph Node Metastasis.

LOO Leave-One-Out.

MALDI Matrix Assisted Laser Desorption Ionisation.

MS Mass Spectrometry.

MS/MS Tandem Mass Spectrometry.

MSI Mass Spectrometry Imaging.

NB Naive Bayes.

PCA Principal Component Analysis.

SNR Signal-to-Noise Ratio.

SVM Support Vector Machine.

TMA Tissue Microarray.

TOF Time-of-Flight.

Abstract

Proteomic imaging mass spectrometry is an emerging field, and produces large amounts of high-dimensional data. We propose approaches to extracting useful information from these data — two of particular note. The Difference in Proportions of Occurrence Statistic (DIPPS) applies to binary data and leads to easily interpretable maps useful for exploratory analyses and automated generation of feature lists that can be used to standardise comparisons between datasets. The second approach, based on Canonical Correlation Analysis (CCA), reduces the high-dimensional data to features strongly related to classes and leads to good classification. Applications to cancer data show the success of these approaches.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed:

Date:

Acknowledgements

First, my thanks go to my primary supervisor, Associate Professor Inge Koch. Inge has consistently challenged me when I needed to be challenged, and been helpful when I needed help. As well as offering valuable expertise, she has also provided guidance and without her occasional course-correcting nudges I certainly would have meandered. In particular, I must commend Inge for her patience during my painfully slow progress towards becoming a better writer and for her immense efforts proof reading my thesis for mathematics, overall structure, and more. Her consistency over the years it took me to complete this work has been incredible. Even when she moved interstate I barely noticed the difference in her availability — if ever I needed to discuss something, she would always have time for me, and for that I am greatly appreciative. I could not have asked for a better supervisor.

My thanks also extend to my co-supervisor Professor Peter Hoffmann. In each case when I needed something from Peter — whether it be an opinion, a signature, or anything else — he was always right there to support me and provide whatever I needed. Peter introduced me into the proteomics community, a field that I had no previous experience in. His experience and comprehensive knowledge of the significant contributors to the field have been invaluable resources. Also, this project has been a collaborative effort with Peter's group, the Adelaide Proteomics Centre (APC), and would never have been possible without their combined help. So I would like to extend my thanks to everyone from the APC.

For the first couple of years I worked closely with Dr. Johan Gustafsson and Dr. Stephan Meding, whose insightful explanations and very high standards of technical expertise were impressive, and extremely helpful during the course of this work. Johan and Stephan, with the help of James Eddes, helped introduce me to some of the foundational ideas in the field, and this solid foundation allowed me to build the rest of this work upon it. In the later years of this work, Dr. Florian Weiland and Dr. Dan Kortschak both helped to keep my waning motivation high by sharing an interest in some of the more technical aspects of my work. Dan helped, and continues to help, introduce me to software engineering, and was a friend during some of the hardest periods of my thesis. I extend my sincere appreciation to each of these people, for all the reasons mentioned above, and more.

For proof reading and correcting this thesis from the biology and mass spectrometry perspectives my thanks also go once again to Peter and Johan. Johan's comments in particular were very thorough, insightful, and helpful in improving this thesis. Additionally, for proof reading this thesis for grammar, Melanie Föll, Mark Witham, Imants Cielens, Gordon Wellman and Peter Cassidy should be commended on their efforts and each have my deep gratitude.

During this work I also rediscovered my love of the outdoors — both hiking and rock climbing — and this helped me maintain a life outside of university. For their contributions to this discovery and their and shared enthusiasm for it, I would like to thank Alex Mackay, Joshua Trestrail, Elisa Chaplin, Peter Arcidiaco, Jason Morton, Garth Wimbush, and the rest of the beautiful Adelaide rock climbing community.

Many more friends deserve my thanks, but I want to thank Stephen Wade and Alison Langsford specifically, both of whom have been particularly good friends during the past few years. Stephen provided a voice of reason, a sympathetic ear, and has remained a good friend long after leaving the university. Along with Josephine Varney, Wei Xian Lim, Paul Tune, and others, he also made the post-graduate room a more enjoyable space to spend time within. For many years, Alison has been a close friend and a person I could always confide in, bounce thoughts against, and I feel I leaned on her far more than could be expected of any ordinary friend. For these things and more, you each have my deep appreciation.

Finally, the one constant in my life — my family: My parents, Saul and Tamar; my brother, Orr, and his more recent additions to our family Aviv and Tevelle. You have all provided me with a source of calm and a sense of home, and for that there are no words.

Introduction

It should be emphasised that although this project is primarily based in the discipline of statistics, or perhaps more accurately bioinformatics, a large component of the work is cross-disciplinary with proteomics, and as such I aspire to represent both points of view to some degree. Proteomics is a rapidly growing area which deals with the identification and characterisation of proteins, most commonly by a so-called 'bottom-up' approach that uses peptides resulting from proteolytic cleavage of the proteins with an enzyme. There are also 'top-down' proteomics approaches that use intact proteins, but I will exclusively consider bottom-up approaches. Many different proteomics methods include Mass Spectrometry (MS) based identification steps. Chapter 1 includes a brief introduction to some of these methods. The focus of this thesis is on one particular, relatively new, application of MS called Matrix Assisted Laser Desorption Ionisation (MALDI)-Mass Spectrometry Imaging (MSI). In contrast to many other more established methods in proteomics, MALDI-MSI has not yet attracted as much attention in the statistics/ bioinformatics literature, although some approaches to the analysis of MALDI-MSI data have been covered in proteomics/ mass spectrometry journals — see Norris et al. (2007); Jones et al. (2012); Gessel et al. (2014); Stone et al. (2012); Alexandrov et al. (2010); Alexandrov and Kobarg (2011); Alexandrov et al. (2013) and references therein. MALDI-MSI can produce large datasets with complicated structure and as such requires the development of novel statistical tools in order to analyse and interpret. The goal of this work is to develop methods that can help in the analysis and interpretation of MALDI-MSI data.

There are two strengths of MALDI-MSI that we will focus on, and these two foci split the work in this thesis into two parts:

- MALDI-MSI can preserve spatial information in the data that would otherwise be lost. Taking advantage of this spatial information is the focus of Chapters 2 and 3, where we introduce clustering as an approach to separate spatially distinct regions in an automated fashion. We then also suggest the Difference in Proportions of Occurrence Statistic (DIPPS) which we use in a feature extraction approach to characterising the regions separated by clustering. This DIPPS-feature extraction provides a quick and easy way to identify potentially interesting targets for follow-up experiments. Such an automated approach to identifying targets is useful as the standard approach is to manually consider each feature and this can be time consuming and even biased, particularly when considering multiple large MALDI-MSI datasets.

- Data can be collected from large cohorts of patients through the use of Tissue Microarrays (TMAs) combined with MALDI-MSI. Having access to data from many patients allows for classification problems of diagnostic relevance to be addressed. Taking advantage of the ability to collect data from large patient cohorts is the focus of Chapters 4 and 5, where we consider different approaches to the classification of MALDI-MSI data from TMAs. One of the

more promising approaches we discuss is that of Canonical Correlation Analysis (CCA)-based variable selection, which not only seems to perform well in the classification context, but as a by-product also identifies key peptides which can be further investigated in follow-up biomarker validation studies.

In each of these two parts, we first introduce and discuss the methods we propose to use (in Chapters 2 and 4 respectively), and then demonstrate the use of these methods on real data in more detail and discuss the biological relevance of the results (in Chapters 3 and 5 respectively).

The research carried out for this thesis has been published in four papers. Two papers, Winderbaum et al. (2015) and Winderbaum et al. (2016), are method-focused publications and correspond to the two foci above. The discussion in Chapters 2 and 3 overlap with the ideas of Winderbaum et al. (2015) and explores these ideas in more detail. Similarly, the main conclusions of the discussion in Chapters 4 and 5 are summarised in Winderbaum et al. (2016). The other two papers, Gustafsson et al. (2015) and Mittal et al. (2016), are application-focused and correspond to the glycan and endometrial cancer applications introduced in Sections 1.5.3 and 1.5.2 respectively. Application-specific results are presented in Section 3.2.2 and Chapter 5, overlapping with results discussed in each of these two application-focused papers. I have also presented the work, now published as Winderbaum et al. (2015) at several conferences, specifically giving a talk at the Statistical Society of Australia Inc. Young Statisticians Conference in 2013, a talk as an invited speaker to the Statistical Society of Australia Inc. Austrian Statistical Conference in 2014, and presenting a poster at the Australasian Proteomics Society Annual Lorne Proteomics Symposium in 2015.

We have access to several high quality applications of MALDI-MSI data through our collaborative work with the Adelaide Proteomics Centre (APC)[1]. Our collaboration with the APC on several of these applications has also produced a number of tangential publications of which I am not the main author — including Gustafsson et al. (2015) and Mittal et al. (2016). In total we consider three such applications of MALDI-MSI in this thesis:

- Ovarian Cancer — We consider approaches to the exploratory analyses of MALDI-MSI data in depth, and make extensive use of the ovarian cancer data of Gustafsson (2012) to illustrate these methods. In Chapter 2 we demonstrate the separation of cancerous tumour tissue from its surrounding non-tumour tissues by using an automated clustering approach. We then suggest a DIPPS-feature selection scheme for selecting a short-list of peptides that are more highly expressed in tumour tissue than non-tumour tissue. We published this DIPPS-feature selection approach as Winderbaum et al. (2015). In Section 3.1 we consider the results of applying this combined feature extraction approach to many sections from the same and different patients, and comparing the results to explore within and between patient variability in MALDI-MSI data.

- Murine Glycans — It would be of interest to detect glycans with MALDI-MSI, rather than just peptides and proteins. In Section 3.2 we demonstrate that the DIPPS-based feature extraction step developed with the ovarian cancer data in mind can also be used to quickly and easily produce a short-list of

[1] `http://www.adelaide.edu.au/mbs/proteomics/`
Level 1, Molecular Life Sciences
The University of Adelaide
SA 5005 Australia

potential glycans for validation in the MALDI-MSI data. This experiment successfully demonstrated that glycans can in fact be detected using MALDI-MSI (Gustafsson et al., 2015), and this opens up new applications in cancer research.

- Endometrial Cancer — Taking advantage of the second strength of MALDI-MSI noted above, Mittal et al. (2016) collected data from the primary tumours of a cohort of endometrial cancer patients using two TMAs. If it is possible to predict these patients' Lymph Node Metastasis (LNM) status from these primary tumour data, this would give the surgeon a diagnostic tool to decide if the lymph nodes need to be removed. Removing the lymph nodes is associated with serious complications for the patient, and LNM is highly relevant to survival and treatment, so this decision is important for producing positive patient outcomes.

  We introduce some classification and variable reduction methods, as well as our novel approach to pre-processing and normalisation of these data in Chapter 4. In Chapter 5 we consider the results of applying these methods to the endometrial cancer MALDI-MSI data, concluding that LNM can be predicted from these data to a significant degree, and therefore warranting further studies applying this method as a diagnostic tool for LNM status in the clinic. We published these results and our approach to this classification problem as Winderbaum et al. (2016).

Each of these applications of MALDI-MSI are introduced in more detail in Section 1.5. Note that throughout this thesis we introduce many established ideas from a number of fields, including proteomics, mass spectrometry, and statistics. Where possible we include specific references to these ideas, but some of the ideas we discuss are commonly known in a particular field. For any such background knowledge without a specific reference included, please see Lovric (2011) for proteomics and mass spectrometry background, Koch (2013) for multivariate statistics, or

It should also be noted that a significant proportion of the work that went into this project involved writing software tools to handle large MALDI-MSI datasets. Although not explicitly discussed in this thesis, all associated code is available from GitHub[2], or at request from the author[3].

[2] URL: `https://github.com/armadilloa16`

[3] email: `lyron.winderbaum@student.adelaide.edu.au`

Chapter 1

Background and Data

In order to be self-contained in terms of required proteomics background knowledge, we begin with a short introduction to the broad concepts of proteomics (Section 1.1) and Mass Spectrometry (MS) (Section 1.2). In Section 1.3 we introduce some more specific proteomics topics that will be relevant to our context. We then discuss the technique that is the main focus of this project, Matrix Assisted Laser Desorption Ionisation (MALDI)-Mass Spectrometry Imaging (MSI), in Section 1.4. Finally, in Section 1.5 we introduce data sourced from three different applications of MALDI-MSI — the study of: peptides from ovarian cancer tumours embedded in surrounding healthy tissues, glycans from murine kidney tissue, and peptides from endometrial and vulvar cancer tumour tissues arranged in Tissue Microarrays (TMAs). Each of these applications highlights different aspects of the potential in MALDI-MSI data. The remainder of this thesis is concerned with developing data analysis methods for such applications. For example, clustering of the ovarian cancer data highlights the ability of MALDI-MSI to separate tissue types spatially. Classification of the endometrial data highlights the ability of MALDI-MSI to contribute to diagnostics by use of TMAs.

The discussion of proteomics in this chapter is biased towards points relevant to MALDI-MSI, as this is our main interest. For more complete reviews of proteomics as a whole and the role of MALDI-MSI, see Mallick and Kuster (2010), Schwamborn and Caprioli (2010) and references therein. Wu et al. (2003) also provide an overview of some of the statistical challenges inherent in working with proteomics MS data.

1.1 Proteomics

Proteins are biological molecules involved in most cellular processes and consisting of a sequence of amino acids that are chemically connected by 'peptide bonds'. When connected in this fashion, each amino acid in a protein is referred to as a residue. There exist 22 naturally occurring proteinogenic amino acids in eukaryotes — organisms whose cells contain nuclei. Naively, it could be said that this means there are 22^n possible unique linear proteins of n residues — in reality it is a smaller number than this, but a very large number nonetheless. Each gene can simplistically be thought of as code for a protein's amino acid sequence. Living organisms transcribe a gene and synthesise the corresponding protein as a linear amino acid sequence which then undergoes post-translational modification and folds into a complex three dimensional structure w hich usually determines the protein's function and intra-cellular location. The compounding diversity of: linear amino acid sequence post-translational modification, and three dimensional folding is what allows proteins to fill such a wide variety of functions. The complete set of proteins

which exist in a given cell, tissue or biological fluid, under defined conditions, is termed its proteome (Wilkins et al., 1996). Proteomics is the study of proteomes, and often of how proteomes change. Proteomes vary considerably between different cellular states and understanding these variations can provide remarkable insights. A cell in the lining of your stomache would have a virtually identical genome to a neuron in your brain, but they would contain dramatically different proteomes, and the differences between their proteomes is what allows for such dramatic differences in their phenotypes — behaviour and functions. We are particularly interested in the effects diseases have on the proteomes of cells. Studying these effects can elucidate mechanisms involved in disease behaviour and provide insight into the development and progression of diseases. Such fundamental knowledge could lead to new approaches for both diagnosis and treatment (Casadonte and Caprioli, 2011).

The study of proteomics includes the identification, quantification, and/or localisation of proteins in a sample (Ong and Mann, 2005). Samples are often sourced from a cell culture, but other biological sources are possible — such as: saliva (Vitorino et al., 2004), blood (Liotta et al., 2003; Thadikkaran et al., 2005), tissue (Chaurand et al., 2004), and even animal specimens (Khatib-Shahidi et al., 2006). Proteomics is an extremely broad field, but techniques that can characterise the proteins present in a sample are needed across almost all areas of the field. MS is the core method in proteomics for characterising and identifying the proteins in a sample (Yates et al., 2009). Many different technologies have been developed in proteomics MS (Aebersold and Mann, 2003). In the following section we explain the principles behind MS and its role in various applications within proteomics, and in particular how it is extended for use in MALDI-MSI.

1.2 Mass Spectrometry

A mass spectrometer measures the Mass-to-Charge Ratio (m/z) ratio of molecules ionised from a given sample. To achieve this, a mass spectrometer invariably consists of three crucial components:

1. An ion source,
2. A mass analyser, and
3. A detector.

As its name implies, the ion source converts sample molecules into gaseous ions, which allows them to be manipulated using electromagnetic fields. The mass analyser separates the ions according to their m/z by controlled application of electromagnetic fields. The detector counts ions after they have been separated in the mass analyser, thereby producing a mass spectrum of ion counts versus m/z. This section briefly discusses the ion source and mass analyser that we use in imaging experiments, MALDI and Time-of-Flight (TOF) respectively. For MSI on biological samples, MALDI is used almost exclusively. MALDI can also be used in Liquid Chromatography (LC)-MS analyses of proteins and peptides, but Electrospray Ionisation (ESI) is the predominant ion source for LC-MS. LC-MS is very useful in proteomics due to its ability to reproducibly identify the peptide sequences of analytes in complex mixtures. As LC is a core technique in proteomics, a brief introduction to LC, specifically LC-ESI, is included in Section 1.2.3.

Although we peripherally refer to some LC-MS identification results in Section 3.2.2 and Section 5.4, the focus of this work is on MALDI-MSI, and so the

discussion of MALDI-MSI in Section 1.4 is appropriately more detailed than the discussion of LC-ESI-MS in Section 1.2.3. Note that when we discuss LC-MS what we are ultimately referring to is Tandem Mass Spectrometry (MS/MS), which is the fragmentation of analytes to determine their identity. Although the distinction betwen MS and MS/MS is important, it is of only tangential relevance to our work and so for brevity we avoid a detailed discussion and simply refer to LC-MS.

Of course there are other ion source options, often designed to ionise particular types of target molecule. For example, Laser Ablation Inductively Coupled Plasma (LA-ICP) is a popular ionisation method for measuring trace element concentrations such as copper, gold, silver, arsenic, etc. LA-ICP-MS can even be used for imaging, and has been used in conjunction with Gel Electrophoresis (GE) and MALDI-MS to study phospho- and metal-containing proteins (Becker et al., 2009, 2010). LA-ICP-MS is one of the most extreme examples of what are called 'hard' ionisation methods. Hard ionisation methods often fragment large molecules during the ionisation process, destroying information about the original intact analytes — LA-ICP-MS typically fragments analytes down to their elemental composition. The 'soft' ionisation methods such as MALDI and ESI are by far the most popular for mass spectrometric analysis in proteomics as they allow for large biological molecules to be converted into gaseous ions with minimal fragmentation.

1.2.1 MALDI

There are many types of ion source, and most are optimised for particular substances or molecules. MALDI is an ionisation method which avoids significant fragmentation of large organic analytes due to the relatively low energy levels that it operates at. MALDI functions by adding a matrix to the sample, which is a small molecule and acts as the primary absorber for a laser system which provides the ionisation energy. There are numerous matrices available, each with advantages and disadvantages, but in each case the matrix is a small molecule whose role is to absorb at the wavelength of the laser system and transfer the absorbed energy to the sample in a controlled manner, ionising the analytes in such a way that their covalent bonds are not broken. MALDI has the added advantage of almost always producing singly charged ions ($z = 1$) for peptides. The interpretation of MALDI spectra is significantly simplified as $z = 1$ can be assumed for low-mass peptides, meaning the measured m/z can be interpreted as simply molecular mass (m) plus a proton that provides the single positive charge. This reduction in complexity is important in the analysis of complex biological samples, as we will discuss in more depth in Section 1.2.3.

1.2.2 Time-of-Flight (TOF)

Allowing analyte ions to travel through a field-free drift region separates analyte ions by m/z, and this type of mass analyser is coined Time-of-Flight (TOF) as analytes are separated based on the time they spend in the field-free drift region — see Figure 1.1. Other types of mass analyser are also used, but we will focus on TOF-MS as this is the most common approach for MSI, and all the data we will consider in detail were collected using this mass analyser type. In TOF-MS, the sample is ionised (given a charge z) before being accelerated through a potential difference, V, to acquire a fixed amount of kinetic energy, $E_k = zV$. The accelerated sample is then allowed to drift a fixed distance, d, through a field-free region to a detector where the TOF can be measured. Because each analyte molecule was given the same kinetic energy, their velocities in the field-free drift region will be determined

by their mass through the relationship for kinetic energy

$$E_k = \frac{1}{2}mv^2. \tag{1.1}$$

Equation 1.1 shows that for constant kinetic energy, E_k, the squared velocity (v^2) of each analyte molecule is inversely proportional to mass (m). If we let the time spent in the field-free drift region (TOF) be t, then $v = \frac{d}{t}$ and as V and d are known constants, we see that mass-to-charge ratio (m/z) is a quadratic function of TOF t,

$$\frac{m}{z} = \frac{2t^2V}{d^2}. \tag{1.2}$$

Equation 1.2 shows how m/z can effectively be measured by flight time — hence Time-of-Flight (TOF)-MS. This whole process, from ionisation to detection, is illustrated in Figure 1.1.

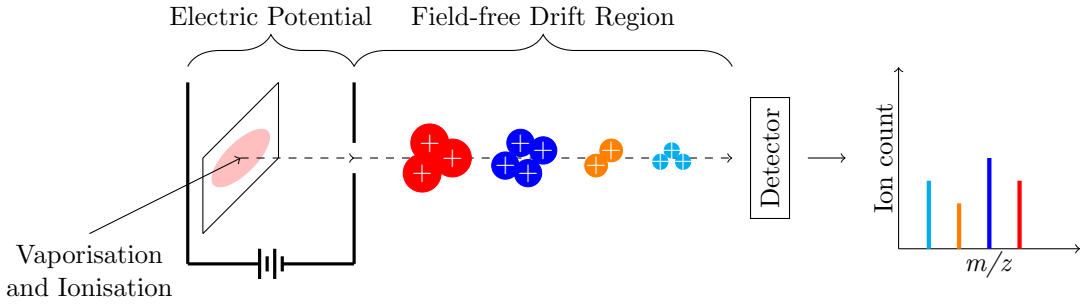


Figure 1.1: Schematic describing the acquisition of a mass spectrum in a linear mass spectrometer. From left to right, sample is ionised, accelerated through an electric potential, allowed to drift through a field free region, and the time of arrival at the detector is measured. Ion arrival time at the detector is transformed into m/z which yields a mass spectrum.

The system described above and illustrated in Figure 1.1 is a simplification when considering most modern instruments. Specifically, when using MALDI, the initial laser pulse results in a small explosion of ions. This explosion results in the analyte ions having variable initial positions and velocities (kinetic energies), and these variations limit the accuracy of the mass spectrometer. A number of improvements to the system described above have been made that can compensate for these small differences in initial position and energy of analyte ions, increasing the accuracy of modern mass spectrometers. Some notable examples of such improvements include:

- **Delayed Extraction:** Variation in the initial positions and energies of analyte ions can be further compensated for by a process called *delayed extraction* (Vestal et al., 1995). Delayed extraction involves incorporating a short delay time between the ionising laser pulse and switching on the accelerating electric potential illustrated in Figure 1.1. This delay allows ions with higher initial energy to drift further into the mass spectrometer than those with lower initial energy before the electric potential is applied. The fact that ions with higher initial energy have travelled further into the mass spectrometer when the electric potential is applied means they traverse less distance within the electric field and are given less energy. Conversely, ions with lower initial energy traverse more distance through the electric field and are thereby given more energy. This process helps compensate for variations in initial energies

and results in ions leaving the electric potential region and entering the field-free drift region of Figure 1.1 with a more consistent total amount of kinetic energy — thereby improving the accuracy of the resulting measurements.

- **Reflectron Mode:** The system described above and in Figure 1.1 is that of a TOF mass spectrometer in *linear mode*, so-called because the analyte ions follow a path through the mass spectrometer described by a straight line. Most modern TOF mass spectrometers still have an option to be operated in linear mode, typically for use in analysing intact proteins, but in practice these instruments are almost always operated in reflectron mode. In reflectron mode, after initial acceleration ions are 'reflected' by a constant electric field at an angle to their initial velocity, causing them to follow a parabolic path and thereby creating a focal point that can compensate for small deviations in initial energy and position of analyte ions (Boesl et al., 1992; Kaufmann et al., 1993). This reflection step essentially replaces the field-free drift region of Figure 1.1. In addition to compensating for some variability in initial energies and positions, reflectron mode allows for a longer effective drift distance d, and these two factors combined significantly improve the accuracy with which the m/z of analyte ions can be measured, and thereby the resolution ultimately achieved.

All the data presented in this thesis were collected in reflectron mode and using delayed extraction.

1.2.3 Fractionation

Proteomics samples are often highly complex — containing many thousands of proteins. Strong signals due to abundant proteins can obscure the weaker signals of less abundant proteins, making these weaker signals difficult or impossible to detect. Also, the m/z values of multiple proteins can overlap, making identification difficult. These complications can be addressed by fractionation — separating analytes in the sample by some physical or chemical property prior to MS acquisition. Fractionation can be done at either the protein level, or at the peptide level, corresponding to fractionation before or after proteolytic cleavage respectively. Fractionation has become standard in proteomics, see Wasinger et al. (1995). GE and LC are examples of common fractionation techniques in proteomics, with LC being the predominant fractionation technique (Gygi et al., 2000; Rogowska-Wrzesinska et al., 2013). For an overview and review of recent approaches in LC-MS, see America and Cordewener (2008). Although we only tangentially incorporate results that involve fractionation in this work, the results we do present are from LC-MS at the peptide level.

GE typically fractionates on the protein level, and separates analytes by size, charge, or both — as is the case in 2D-GE, a very popular method in proteomics. This is achieved by using an electric field to force molecules through a gel, often polyacrylamide, that acts as a 'sieve'. Smaller molecules find that their movement is restricted less by the gel media than larger molecules, and in this way molecules are separated by size.

LC makes use of columns to bind molecules to a stationary phase — an immobile surface with gaps to allow solvent to be pushed through. The molecules are subsequently eluted over time using a changing gradient of mobile phase solvent. Traditionally the stationary phase was hydrophilic, but 'reverse-phase' chromatography is popular in proteomics, in which the stationary phase is hydrophobic and the hydrophobicity of the mobile phase is increased gradually with time. Coupling

this fractionation method directly to a mass spectrometer (LC-MS) allows molecules to be characterised by MS as they elute off the column one by one, reducing the complexity of any single spectrum tremendously.

Both LC-MS and 2D-GE are very powerful methods in terms of their ability to characterise and identify proteins in a sample. However both these methods involve the homogenisation of a sample prior to analysis, and this step destroys any information about the spatial distributions of proteins within a sample. The loss of spatial information in these established methods motivated the development of direct tissue analysis using MALDI-MSI (Cornett et al., 2007; Groseclose et al., 2008; Gustafsson et al., 2011).

1.3 Histopathology and Formalin Fixation

We are interested in the use of MALDI-MSI to detect spatial patterns in tissue and so it is important to consider other histopathology techniques. Immunohistochemistry (IHC) is the established 'gold standard' for mapping the spatial distribution of a protein, and so any discussion of MALDI-MSI in the context of mapping such spatial distributions requires comparison to, and validation by, IHC. In Section 1.3.1 we briefly introduce histopathology and IHC in comparison to MALDI-MSI. Ultimately, we discuss how MALDI-MSI can complement the use of these methods.

There are different methods for preserving or 'fixing' the spatial information in the tissue for analysis. In the field of MALDI-MSI, fresh frozen tissue samples have historically been used, but recent advances have allowed for the use of Formalin Fixed and Paraffin Embedded (FFPE) tissue. This has had widespread impact on the field as FFPE tissue samples can be stored long term relatively cheaply, and so large archives of FFPE tissue samples exist. Access to larger sample sizes via these FFPE tissue archives has allowed for previously difficult research questions to be tackled, as we discuss further in Section 1.4.2. In Section 1.3.2 we introduce the concepts underlying fresh frozen and FFPE tissue samples, including a brief comparison of their respective advantages and a discussion of the methodological developments that have led to MALDI-MSI analysis of FFPE tissue. All the applications of MALDI-MSI we consider in this thesis are on FFPE tissue.

1.3.1 Histopathology and Immunohistochemistry

Histopathology is the study of the anatomy of tissues at the microscopic scale, and is typically performed by a pathologist, who can often provide diagnostic information on a disease by examination of a stained section of tissue using a light microscope. The staining step is important as tissue has little inherent contrast and, as well as adding contrast, staining can highlight features of interest. Hematoxylin and Eosin (H&E) is one of the most common stains used in histopathology, highlighting cellular nuclei and cytoplasm respectively. We use images of H&E stained tissue as a baseline of spatial structure to compare the results of many of our analyses. In some datasets, we also use annotations made by a pathologist on the basis of H&E stained tissue.

IHC refers to the use of an antibody to stain tissue in an attempt to highlight the corresponding antigen (protein). Given an antibody for the protein of interest a stain is usually achieved either by conjugating the antibody to an enzyme, such as peroxidase, that can catalyse a visible reaction, or by tagging the antibody with a fluorophore. Regardless of how the antibody is made visible, this process allows for

the spatial distribution of a particular protein to be visualised, given a sufficiently specific antibody.

In IHC the protein of interest and associated antibody are chosen *a priori*, and as such this method cannot discover previously unidentified proteins with interesting spatial distributions. MALDI-MSI does not have this limitation as it does not target specific proteins — and so has the capacity to lead to the discovery of new biomarkers (Schwamborn and Caprioli, 2010). However MALDI-MSI has only a limited ability to identify proteins, so even if a mass is found to have an interesting spatial distribution, follow-up identification and validation experiments are required. Potential identifications for masses of interest can be inferred by mass-matching to peptides identified in parallel LC-MS experiments on similar tissue, as demonstrated by Meding et al. (2012). Identification is one of the strengths of LC-MS, but LC-MS provides no information on distribution within the tissue. As IHC is cheaper, easier, and much more established than MALDI-MSI, it makes sense to find masses with interesting spatial distributions by MALDI-MSI, infer parent proteins by LC-MS, and validate the spatial distributions of these proteins by IHC.

1.3.2 FFPE vs Fresh-Frozen

In this section we briefly discuss the differences between FFPE and fresh-frozen tissue. FFPE tissue is prepared by first immersing tissue in a formalin solution and then embedding the tissue in paraffin for storage. Formaldehyde in the formalin solution creates covalent cross-linking bonds between proteins thereby 'fixing' the tissue by interrupting biochemical reactions, preventing decay, and causing the tissue structure to stabilise (Fox et al., 1985). This fixation process can be partially reversed using heat in the presence of excess water, and this process is often called 'antigen retrieval'. FFPE samples can be stored indefinitely at room temperature while proteins (and even nucleic acids) are still recoverable for detection many years after fixation — making FFPE samples a crucial resource in retrospective or large-sample studies. Fresh-frozen tissue is prepared by rapidly reducing the temperature of the tissue. This is typically achieved by placing the tissue sample in liquid nitrogen. Care must be taken to preserve spatial information during the cooling process (Schwartz et al., 2003).

FFPE tissue can be stored at room temperature, while fresh frozen tissue requires expensive refrigeration, and so FFPE tissue is much cheaper and easier to store for long periods of time. For this reason, FFPE tissue is the international gold standard for tissue sample storage, and large archives of FFPE tissue exist, often with complete patient history and meta-data (Hood et al., 2005). It is usually not feasible to obtain such a large number of fresh frozen samples, so when designing an experiment for which a large number of samples will be needed, using FFPE tissue is preferable. Proteomic analysis of FFPE tissue has been difficult due to the cross-linking of the proteins (Hood et al., 2005), and so until recently, fresh frozen tissue has been the most common sample for proteomic analysis (Poschmann et al., 2009). As such, most studies have involved only a small number of patients due to the limitations of using fresh frozen tissue. One of the foci in the applications of MALDI-MSI that we consider in Section 1.5 is the application to relatively large sample sizes, and we will consequently use FFPE tissue.

Early work on antigen retrieval originates from the immunology field, as discussed by Brown (1998). First attempts involved simple enzymatic cleavage, but Shi et al. (1991) introduced the first method to employ additional heating via a microwave source. Since the work of Shi et al. (1991), more heat-based methods were

developed and Shi et al. (2005) discussed the standardisation of these methods as they matured in the IHC field. Eventually these developments were transferred into the increasingly popular mass spectrometric fields, as evidenced in Palmer-Toy et al. (2005) and Crockett et al. (2005). Groseclose et al. (2008) suggested a methodology for antigen retrieval and subsequent MALDI-MSI of FFPE lung tissue. Gustafsson et al. (2010) applied the methodology of Groseclose et al. (2008) to FFPE ovarian cancer tissue, and proposed an improved antigen retrieval methodology. In all the applications we consider in Section 1.5, the methodology of Gustafsson et al. (2010) was used for antigen retrieval.

1.4 MALDI-MSI

MALDI-MSI is a technique which collects a MALDI-MS spectrum from many points on the surface of a tissue sample. As we have already introduced MALDI-MS in Section 1.2, here we will focus on the imaging aspect, and properties of MALDI-MSI — i.e. the process of collecting many MALDI-MS spectra from spatially distributed points across the surface of a tissue sample.

There are two main goals that use of MALDI-MSI can facilitate:

- To resolve the spatial distributions of biomolecules of interest within tissue samples.
- To acquire data from large numbers of patients by use of TMAs.

We discuss advantages and disadvantages of approaches to these two main goals in Section 1.4.1 and Section 1.4.2 respectively. For a review of MALDI-MSI see Seeley and Caprioli (2011). Groseclose et al. (2007) and Aoki et al. (2007) also provide discussions on MALDI-MSI.

Other MSI methods exist, even within proteomics — Becker et al. (2009, 2010) are good examples of this, making use of Laser Ablation Inductively Coupled Plasma (LA-ICP)-MSI to study metalloproteins. We focus specifically on MALDI-MSI due to the usefulness of MALDI in proteomics, as discussed in Section 1.2.1. The process of taking many measurements on the surface of a tissue sample by MALDI-MSI is illustrated in Figure 1.2. In order to collect MALDI-MSI data from a tissue section as depicted in Figure 1.2, a number of sample preparation steps must first be completed:

- Antigen Retrieval - If FFPE tissue is used, cross-linking caused by formalin fixation must be partially reversed prior to analysis, as discussed in Section 1.3.2. We use the method suggested by Gustafsson et al. (2010) for citric acid antigen retrieval.
- Enzyme Digestion - It can be useful to digest proteins by use of an enzyme (eg. trypsin) prior to acquisition, as the smaller peptides produced by enzymatic cleavage can be measured more accurately. All the peptide data we consider was acquired after trypsin digestion, so the signals we observe are tryptic peptides. For the glycan data a different enzyme, PNGase F, was used to cleave off the asparagine (N)-linked glycans of interest.
- Internal Calibrants - Known calibrants are added so that each spectrum will contain calibrant peaks that can be used for mass-calibration. We follow the procedure for internal calibration suggested by Gustafsson et al. (2012). Coincidentally, these internal calibrants can also be useful for data quality control as discussed in Section 4.4.

- Matrix deposition - A matrix, as briefly discussed in Section 1.2.1, needs to be deposited onto the tissue in order to facilitate ionisation. In MALDI-MSI there are two approaches to this, which we introduce and compare in Section 1.4.1.

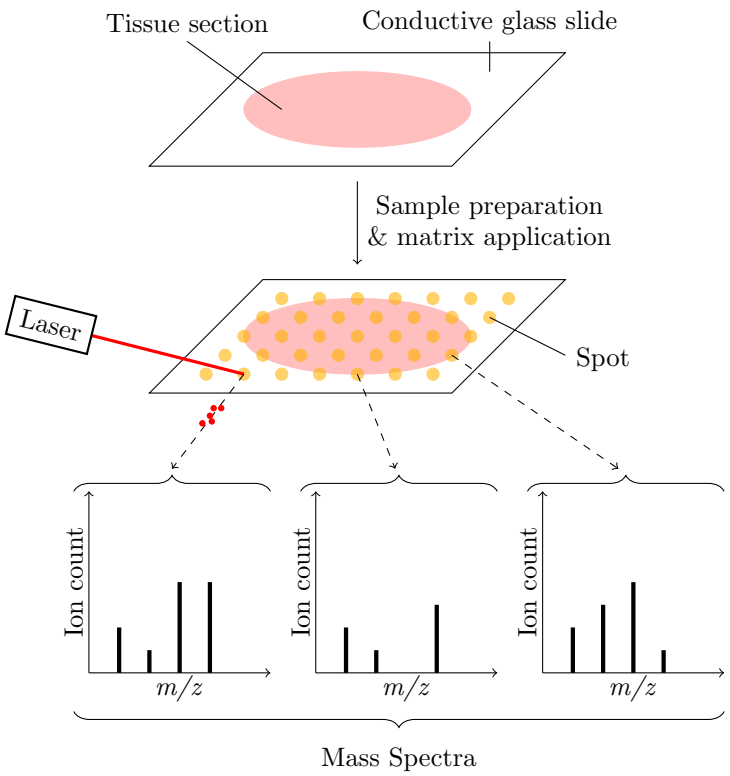


Figure 1.2: A schematic diagram illustrating the work-flow for MALDI-MSI. The depicted spots reflect a spotting approach — if a spraying approach was used the reagent would be uniformly distributed accross the slide. Note that 'Ion count' is often simply called 'intensity'.

1.4.1 Spotting vs. Spraying

Care needs to be taken in all of these sample preparation steps to preserve spatial information by limiting the mobility of molecules of interest on the surface of the tissue (Schwartz et al., 2003). In some of the sample preparation steps, most notably the internal calibrant and matrix deposition steps, limiting the mobility of molecules of interest can be achieved by ensuring that reagents are deposited in non-overlapping 'spots'. There are two widely accepted approaches for depositing reagents onto the surface of the tissue, which we will call 'spotting' and 'spraying' respectively. Here I will briefly introduce these two approaches, and discuss their relative advantages and disadvantages. For a more detailed discussion, including some examples of instrumentation for various applications, see Walch et al. (2008).

Spraying

Spraying involves aerosolising the reagents and depositing them as a fine mist. This involves the adjustment of many tuning parameters such as droplet size, concentration, and total volume deposited as well as the number of spray deposition cycles. There are two main disadvantages to spraying:

- Risk of mobilising the molecules of interest — by random chance multiple droplets could merge over a significant region of tissue, allowing molecules of interest to disperse across this region.

- The possibility of gaps — in order to automate data acquisition spectra are collected from arbitrary positions, some of which may, by random chance, not have matrix on them, and thus no molecules will ionise from those positions.

However, the impact of these disadvantages can be controlled and greatly reduced by careful optimisation of the tuning parameters. One of the main advantages of spraying is the ability to push the limits of lateral resolution (centre to centre distance between spectra acquisition locations). All the applications we consider use spraying, with lateral resolutions ranging from $50\,\mu$m to $100\,\mu$m. Guenther et al. (2011) have pushed the limits of this technology, demonstrating that it is possible to resolve images at lateral resolutions as low as $5\,\mu$m under certain conditions.

Spotting

Spotting utilises an instrument comparable to an ink-jet printer to deposit reagent at pre-allocated positions (or spots) across the surface of the tissue. After the antigen retrieval, trypsin, internal calibrants, and matrix would be spotted on the same grid of positions, and a mass spectrum collected from each spot. This guarantees that each spot will contain both internal calibrants and matrix. Spotting has the additional advantage that it guarantees there will be no overlap between spots, thereby restricting the mobility of molecules to within each spot. The disadvantages of spotting when compared to spraying are: it is slower (depositing individual spots is time consuming), the printing instrumentation is more expensive, and the minimum achievable lateral resolution (centre to centre distance between spots) is generally more coarse — typically $\geq 150\,\mu$m.

In addition to the glycan data we consider in Section 1.5.2, Gustafsson et al. (2015) also considered spotted droplet data. Spotting can be most useful when the objective is to acquire a small number of high-quality spectra and a low lateral resolution is not a priority. This often occurs when the aim is to provide a proof-of-principle demonstration of a new methodology or technique, as was the case when Gustafsson et al. (2015) demonstrated that it was possible to use MALDI-MSI to map the spatial distributions of asparagine (N)-linked glycans by using PNGase F to cleave the glycans of interest from their parent proteins. The promising large droplet results then prompted the follow-up experiment using spraying to achieve better lateral resolution, which which we consider in Section 1.5.2.

1.4.2 TMAs for MALDI-MSI

A Tissue Microarray (TMA) is a physical array of cylindrical tissue cores extracted from blocks of preserved patient tissue. Cores are typically less than 2mm in diameter (Meding et al., 2012) and the number of cores in a single TMA block depends on the size of the cores — smaller diameters allow more cores to fit in a single block, but less tissue to be analysed in each core. Using cores $\sim$ 1.5mm in diameter it is typical to fit of the order 50 cores in a single TMA. This arrangement allows for the parallelisation of tissue analyses, as the entire TMA can be treated as a single sample. For example, the entire TMA can be sectioned as a single block, resulting in tissue from many samples being represented in each section — an H&E stain of

such a section is shown in Figure 1.3, including tumour annotations from a pathologist. TMAs were originally developed in order to parallelise IHC staining, but more recently have attracted attention for MALDI-MSI.

Construction of a TMA involves the arrangement of tissue cores from multiple patient samples (biopsies) into a new 'TMA' block that can then be sectioned as one. The application of MALDI-MSI to TMAs allows for data to be collected from a large number of patients very quickly — see Groseclose et al. (2008); Steurer et al. (2013). A one-day sample preparation of sections from such a TMA allows for an overnight MALDI-MSI experiment to collect data on more than 50 cores in one 24 hour period. This approach is significantly faster than more traditional tissue analysis approaches, such as LC-MS. If MALDI-MSI can be shown to produce diagnostically relevant information, its speed would allow for it to potentially be applied for clinical screening of individual samples (Casadonte and Caprioli, 2011).

Figure 1.3: A H&E stained section of a typical TMA, annotated by a pathologist in black to highlight tumour regions. This corresponds to a section of the TMA labelled 'EB' of the endometrial data we introduce in Section 1.5.3 and Table 1.4.

1.5 Data

We consider datasets from three applications of MALDI-MSI:

- Ovarian Cancer — These whole-section data relate to ovarian tumours embedded in peritoneal tissue (elements of the internal abdominal wall) from three patients. Further details can be found in the work of Gustafsson (2012). These data contain spectra from both tumour tissue and surrounding non-tumour tissues excised as a 'margin of error' around the tumours during surgery. These are peptide data and the main interest is to find peptide signals that differentiate between different tissue types, particularly peptides that are present more in tumour than the surrounding non-tumour tissues. We approach this in a two-step process, first using clustering to separate the tissue types in

an automated manner, second using a Difference in Proportions of Occurrence Statistic (DIPPS)-based feature extraction step to find a short-list of potential peptides of interest. We introduce and discuss this two step process in Chapter 2. In Section 3.1 we then compare these short-lists of tumour-identifying peptides between patients, to separate the within-patient variability from the between-patient variability and detect peptides specific to the short-lists from certain patients — i.e. peptides that are over-expressed by tumour cells in particular patients but not others. In the future, such peptides could form a starting point for experiments investigating individualised treatments, but at this initial stage the ability to detect such peptides acts primarily as a proof-of-principle that can then be followed up.

- Murine Glycans — These data relate to sections of murine kidney treated with PNGase F to release asparagine (N)-linked glycans. The object of these data is to serve as a proof-of-principle that glycans can effectively be released *in situ* and measured by MALDI-MSI. In Section 3.2 we demonstrate how the second (DIPPS-based feature extraction) step of the approach introduced in Chapter 2 can be applied in this context to produce a short-list of potential glycan masses.

- TMAs — We consider two TMA applications of MALDI-MSI, one primary application (to endometrial cancer) that we will focus on and a secondary, smaller, application (to vulvar cancer) that we will use primarily to validate and replicate results obtained from the endometrial cancer application. The endometrial cancer data relate to TMAs of endometrial cancer, providing access to data from a relatively large number of patients (more than 40), as discussed in Section 1.4.2. The objective is to demonstrate that these data can be used to predict clinically relevant diagnostic information, justifying further research into the use of MALDI-MSI of TMAs for clinical screening and diagnostics. The diagnostic variable of interest is lymph node metastasis status, which we attempt to classify on the basis of the MALDI-MSI data in Chapter 5. The secondary vulvar cancer data relate to a similar diagnostic variable of interest relating to metastasis, but provide access to data from a smaller number of patients (still more than 20). We will use the vulvar cancer data to investigate whether the conclusions we reach through investigation of the endometrial cancer data can be replicated in an unrelated dataset.

All the data originate from FFPE tissue samples that have been sectioned — cut into thin cross-sections, typically $6 - 10\,\mu$m thick. We will use the word 'dataset' to mean the data collected from one such section of tissue, except where otherwise specified. All the data were acquired by MALDI-MSI at the Adelaide Proteomics Centre (APC), but on two different instruments, and using different parameters. Parameter choices involved in the data acquisition are shown in Table 1.1, as well as the citations relevant to each application including several papers we have published in the process of pursuing this research.

All the data we will deal with is in 'peaklist' format, meaning that some pre-processing has already been performed to extract the signals of interest from the raw spectra. This pre-processing has been done in proprietary software (flexControl, flexAnalysis, and flexImaging, Bruker Daltonik, `http://www.bruker.com`), and involves a number of steps: smoothing (Gaussian kernels), baseline reduction (TopHat), and finally peak picking (SNAP). The SNAP algorithm isolates mono-isotopic peaks and defines significant peaks as those peaks with a Signal-to-Noise

Table 1.1: Data Acquisition Parameters

Application	m/z range	Lateral Resolution	Citation
Ovarian Sections	$1000 - 4500$	$100\,\mu$m	Gustafsson (2012) Winderbaum et al. (2015)
Murine Glycans	$800 - 4500$	$100\,\mu$m	Gustafsson et al. (2015)
Endometrial TMAs	$800 - 4000$	$60\,\mu$m	Mittal et al. (2016) Winderbaum et al. (2016)
Vulvar TMAs	$800 - 4000$	$60\,\mu$m	

Ratio (SNR) of two or higher. These pre-processing and peak-picking methods could be improved, and as all our analysis is downstream of the peak-picking (that is, it occurs after peak-picking), any improvements to these pre-processing and peak-picking steps could clearly carry through to our results. However, the object of most of our work is proof-of-principle and so, although interesting, optimisation of pre-processing methods falls beyond the scope of this work.

In Sections 1.5.1, 1.5.2 and 1.5.3 we introduce and discuss details for the three applications of MALDI-MSI respectively. When introducing the details of these applications, we also discuss the objectives of each application, and the statistical approaches we will use in order to address these objectives, in general terms. In Section 1.6 we introduce some statistics background, in particular introducing the terms 'clustering' and 'classification', giving some added context and expanding upon the approaches briefly mentioned here in Section 1.5.

1.5.1 Ovarian Cancer Application

Winderbaum et al. (2015) motivate the study of ovarian cancer by proteomics:

> "Ovarian cancers are virtually asymptomatic and as a result the vast majority of cases are detected when the disease has metastasised. For these patients, radical surgery and chemotherapy are often insufficient to address the disease adequately and many patients relapse. The combination of late-stage diagnosis and unsuccessful treatments makes ovarian cancer the most lethal gynaecological cancer, with advanced stage patients exhibiting a five year survival rate of less than 30% (Ricciardelli and Oehler, 2009; Jemal et al., 2011). The keys to addressing ovarian cancer will be: increasing our understanding of the mechanisms driving cancer progression, identifying molecular markers which can predict treatment success and identifying new treatment targets. As proteins are key functional components of cells and tissues, determining protein distributions in cancer tissue represents a crucial step in addressing these key aims."

Ovarian cancers are known to be quite heterogeneous tissues (Deininger et al., 2008), and this motivates the use of MALDI-MSI in acquiring spatial information that can de-convolute the inherent heterogeneity of the tissue (Gorzolka and Walch,

Table 1.2: Total number of peaks, spectra, and empty spectra (spectra with no peaks) for each of the ovarian cancer datasets.

Dataset Name	# Peaks	# Spectra	# Empty Spectra
A1	1721862	13916	4
A2	1616042	14225	2
A3	1301720	14059	9
A4	1608226	15386	25
B1	993622	8554	0
B2	1201711	11322	0
B3	976379	9253	0
B4	1209893	11018	1
C1	630973	6731	19
C2	727795	9059	119
C3	886368	9419	14
C4	423993	8404	99

Table 1.3: Two peptide m/z values found in many of the ovarian cancer MALDI-MSI datasets and their inferred parent proteins. Peptide sequences and parent proteins were inferred by mass matching to concurrent LC-MS analyses and validated by both *in situ* tandem MALDI-MS and IHC. IHC stains used for validation are shown in Figure 1.4

LC-MS/MS mass $[M+H]^+$	UniProtKB/SwissProt Database Entry Name	Protein Name
1628.8015	ROA1_HUMAN	Heterogeneous nuclear ribonucleoprotein A1
2854.3884	K1C18_HUMAN	Keratin 18

2014). Gustafsson (2012) describes the acquisition of the data we have from a number of embedded ovarian tumours — containing not only tumour tissue, but also a margin of surrounding non-tumour tissues. One of our primary aims in considering these data is to separate the different tissue types, which we do by means of the k-means clustering. Further to separating spectra from different tissue types, we are interested in the extraction of features or variables specific to the tumour tissue and less prevalent in the non-tumour tissues. We present a method for this feature extraction based on ranking the variables by a statistic which we call DIPPS, and then selecting the highly-ranked variables. The process we propose for these clustering and feature selection steps is described in detail in Chapter 2. Another aim when considering these ovarian cancer data is to compare these highly ranked variables across multiple datasets from both the same and different patients, in an attempt to provide a proof-of-principle that differences between patients can be found using MALDI-MSI in this way, potentially providing reasons to pursue future cohort studies making use of TMAs to predict treatment response. We consider the results of our combined clustering and feature selection method on several datasets from the same and different patients in detail in Section 3.1. Chapter 2 and Section 3.1 represent a more comprehensive view of the results published by Winderbaum et al. (2015), and provide us the opportunity to discuss these results in more detail.

The ovarian cancer datasets originate from patients who were diagnosed with

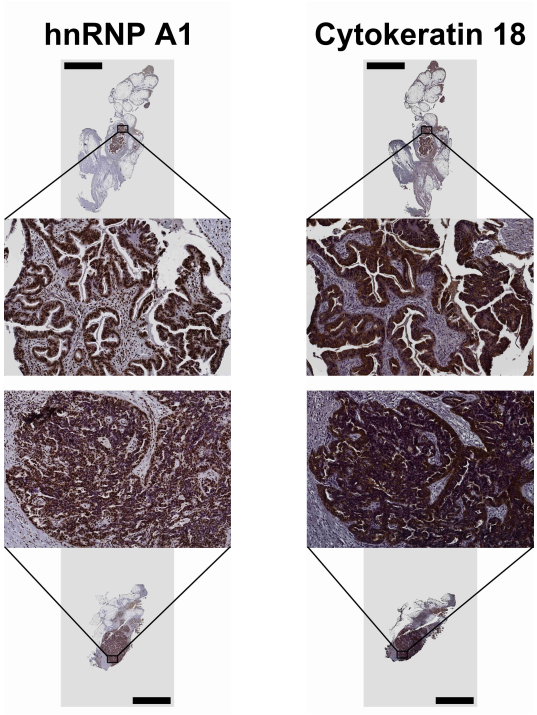


Figure 1.4: IHC stains used for validation of the proteins in Table 1.3 on two of the ovarian cancer patients. Columns correspond to protein antibody stains, rows to patients.

serous ovarian cancer and went through surgery to have tumours removed. We consider 12 datasets collected from sections of surgically excised ovarian cancer tumours — 4 from each of 3 patients, where the 4 datasets from each patient are from multiple tissue sections of the same tumour or tissue block. We will refer to the patients as A, B, and C, and will refer to the four datasets from patient A as; A1, A2, A3, and A4, and similarly for patients B and C. Table 1.2 provides some details for each of these datasets, indicating the scale of these data. The number of empty spectra, that is spectra with no peaks, can be used as a heuristic for quality control as every spectra, even off-tissue spectra, should contain at least the internal calibrants and so a spectrum being empty most likely indicates acquisition conditions that require examination or improvement — inconsistent matrix crystallisation, for example. The number of empty spectra shown in Table 1.2 are generally very small relative to the total number of spectra, indicating no obvious problem with the data acquisition, although amongst all 12 datasets C2 and C4 seem to have notably more empty spectra than others. There is no immediately obvious reason why these two datasets should be any different to the others, but we will notice these empty spectra in further analyses in Chapter 2.

Similar to the methodology of Meding et al. (2012), LC-MS data was collected in parallel for these ovarian cancer samples. Peptide identities of masses in the MALDI-MSI data can be inferred by matching to the masses of peptides identified in the LC-MS data. Two such peptide masses that we find to be of particular interest later in Chapter 2, with their parent proteins, are shown in Table 1.3. IHC validation was also carried out for three selected proteins, including the two of Table 1.3, and the IHC stains for these are shown in Figure 1.4.

1.5.2 N-Glycan Application (in Murine Kidney)

Gustafsson et al. (2015) motivate the study of glycans:

> "The majority of mammalian secretary and membrane proteins are modified through glycosylation: the covalent linkage of polysaccharide moieties (i.e., glycans) to either serine/ threonine (O-linked) or asparagine (N-linked) residues (Pan et al., 2011)." ... "Much of the interest in the analysis of protein glycosylation stems from the observation of altered glycosylation patterns in cancer (Pan et al., 2011). An understanding of these alterations could provide novel biomarkers of disease as well as new treatment targets for anticancer therapies (Abbott et al., 2008)."

The glycan experiment involved the use of an enzyme PNGase F to free N-linked glycans from FFPE tissue in order to be available for analysis, this sample preparation step essentially replaced that of the trypsin digestion discussed in Section 1.4. The objective of the experiment was to demonstrate the proof-of-principle that N-linked glycans could be detected by MALDI-MSI on FFPE tissue. As this was a proof-of-principle experiment and not yet a study of a clinically relevant disease such as cancer (which is intended to be the next application of this method), non-precious murine (rodent) kidney tissue was used. Gustafsson et al. (2015) carried out three different analyses on these samples: LC-MS of homogenised tissue treated with PNGase F, and both large droplet *in situ* MALDI-MS (as mentioned in Section 1.4.1), and high-resolution MALDI-MSI on both tissue treated with PNGase F, and (untreated) control tissue.

For the high-resolution MALDI-MSI data we will consider, spectra were collected from two comparable regions of tissue, one of which was treated with the enzyme PNGase F and the other, 'control', region was not treated with the enzyme. Masses observed in the first region but not in the second are expected to be glycan masses, and further evidence for their identity as glycans was established by Gustafsson et al. (2015) using the additional LC-MS and large droplet data. As this experiment essentially consists of only a single MALDI-MSI dataset, some of the decisions made in preprocessing the ovarian cancer datasets (where a major objective was to compare multiple datasets) are no longer easily justified for the glycan dataset. Specifically, the preprocessing decisions relate to the method for discretisation of the m/z domain, for the ovarian cancer datasets we propose a naive 'data-independent' binning approach (explicitly defined in Appendix A), as this allows for multiple datasets to be combined and compared intuitively. However when this is not necessary, better approaches exist, and in Section 3.2 we discuss this point in more detail, and introduce alternative 'data-dependent' discretisation methods appropriate for the glycan data. Ultimately we apply the 'DIPPS' feature extraction approach introduced in Chapter 2 to select likely glycan mass candidates in the MALDI-MSI data, and match them by mass to identified glycans from the LC-MS as shown in Table 3.2.

1.5.3 TMA Applications

Endometrial Cancer TMAs

As mentioned in Section 1.4.2, an exciting application of MALDI-MSI is to TMAs — as this allows data from large cohorts of patients to be collected rapidly, potentially allowing difficult diagnostic and prognostic problems to be addressed. The problem we will consider is that of diagnosing Lymph Node Metastasis (LNM) in patients with endometrial cancer.

Winderbaum et al. (2016) motivate the classification of LNM in endometrial cancer:

> "Endometrial cancer is the most common gynaecological malignancy in Australia with 2256 diagnosed cases in 2010 and 381 associated deaths in 2011 AIH (2012). The presence or absence of Lymph Node Metastasis (LNM) is the most important prognostic factor in endometrial cancer as patients with localised disease have a 5 year survival rate of 96%, which drops to just 17% for patients with metastatic disease Rungruang and Olawaiye (2012). Accurately staging endometrial cancer is difficult and a large percentage of patients are misclassified prior to treatment Jacques et al. (1998). Although the presence of LNM is confirmed in only around 15% of cases Morrow et al. (1991); Creasman et al. (2006), the majority of endometrial cancer patients undergo radical treatment including the removal of pelvic lymph nodes as a precautionary measure to compensate for our current inability to accurately stage the disease. Lymph node removal is associated with significant complications including lower extremity lymphoedema, which has been described in up to 38% of patients Todo et al. (2010). A classification system based around predictive tissue markers of metastasis would greatly benefit stage I endometrial cancer patients by helping determine optimal treatment strategies that avoid unnecessary, invasive procedures."

LNM is highly predictive of survival but can be difficult to diagnose pre-surgery. Clinical variables such as grade, size, and depth of myometrium invasion are sometimes available pre-surgery but do not correlate well enough with LNM to allow accurate prediction — see Table 1.5. We aim to demonstrate that it is possible to improve on this prediction by applying suitable classification methods to MALDI-MSI data from TMAs.

If a patient is LNM positive and they do not have their lymph nodes excised with their primary tumour the cancer is very likely to recur. Generally the lymph nodes are excised at the same time as the primary tumour, as it is difficult to diagnose the LNM status pre-surgery and the risk of leaving any potentially LNM positive lymph nodes outweighs the additional invasiveness of the surgery to remove them. In addition to LNM status potentially providing more accurate prognostic prediction due to its importance to survival, being able to diagnose LNM status could potentially contribute to the individualisation of patient treatments. Patients who are negative for LNM could be subjected to less invasive surgery as their lymph nodes would not need to be removed. Reduction in the invasiveness of the surgery could then have positive carry-on effects on post-surgery recovery time and quality of life.

The endometrial cancer datasets we consider originate from 57 patients diagnosed with endometrial cancer who had been through surgery to have their tumours excised. Two cylindrical 1.5mm diameter tissue cores from each patients primary tumour were taken and the cores arranged into two TMA blocks — each TMA block consisting of a 7×9 grid of cores, including some control cores and some empty positions. These TMA blocks were then sectioned. We consider 4 datasets, two sections of each of the two TMA blocks. We call the two TMA blocks 'EA' and 'EB' respectively. We will denote the two datasets from EA EA1 and EA2, and similarly denote the two datasets from EB EB1 and EB2. Although each core is taken from primary tumour tissue, some sections of some cores still have non-tumour tissue or are entirely missing due to the three dimensional nature of the cores and the heterogeneity

of the tumours. In order to address this potential tissue heterogeneity each section was also H&E stained and the tumour tissue was manually annotated by a pathologist. Similarly to Table 1.2, Table 1.4 provides some details for these datasets. The additional column '# Cancer Spectra' gives the total number of spectra from regions annotated as tumour tissue by the pathologist. Notice that the total number of cancer spectra is significantly smaller than the number of spectra total — this is likely because of both small margins of off-tissue regions being acquired, and due to the significant heterogeneity of the tissue cores.

Table 1.4: Basic statistics on endometrial TMA datasets

Dataset Name	# Peaks	# Spectra	# Empty Spectra	# Cancer Spectra
EA1	3153313	32562	34	11431
EA2	2799700	26529	54	10122
EB1	3098311	34127	1	12464
EB2	3362057	30343	23	11860

In the ovarian cancer data we refer to datasets as they conveniently separate the data into biological and technical replicates. However in the endometrial cancer data the goal is to make predictions about patients, and many patients are represented in each dataset. We combine all four of the datasets described in Table 1.4, and in our analysis we will partition these data by patient. Also, after closer inspection of the clinical records it was determined that several patients represented in the TMAs are not comparable to the others for the purposes of predicting LNM — belonging to a different sub-type of cancer expected to have very different molecular composition and behaviour. Data from these patients will be completely ignored, and after having removed these inappropriate patients we are left with data from 43 patients, 16 of which are LNM positive. Table 1.5 shows some clinical variables of interest as well as details similar to Table 1.2 and Table 1.4 for these 43 patients.

We will approach the question of predicting LNM on the basis of the MALDI-MSI data using the framework of classification. In Chapter 4 we introduce a variety of classification methods from the literature, some established, and one quite recent. We also discuss the difficulties in classification of such high-dimensional data, and present two different approaches to dimension reduction, the comparison of which is the focus of Winderbaum et al. (2016). We also discuss our approach to pre-processing these data, and our novel approach to what we call 'normalisation' that could potentially improve classification results. After having introduced all these methods in Chapter 4, we apply them to the endometrial cancer data in Chapter 5, and discuss the results in detail. Finally we conclude that useful information that can be used to improve diagnosis of LNM status can be obtained from these data in a variety of ways. The variable reduction method suggested in Winderbaum et al. (2016) works by ranking variables and selecting the highly-ranked ones for use in classification. This approach has the additional advantage that highly-ranked variables can be further investigated as potential biomarkers for LNM, and we consider some promising variables with which to begin this process in Section 5.4.

Vulvar Cancer TMAs

Vulvar cancer is another gynaecological cancer for which, similar to endometrial cancer, LNM is highly predictive of prognosis and important to the choice of treatment.

In Chapter 5 we compare different approaches to the classification of the endometrial cancer data described above. Ultimately we note that certain approaches tend

Table 1.5: Basic statistics on endometrial TMA data, per patient. Grade, Size, and Myometrium Invasion are clinical variables known to be correlated to LNM. These clinical variables are sometimes known before surgery, and have been considered for use in prediction of LNM status but, as reflected in these data, do not allow for very accurate prediction. Note the presence of some missing values, these correspond to patients whose stored clinical data are incomplete.

Patient #	LNM	Grade	Size	Myometrium Invasion	# Spectra	# Cancer Spectra	# Empty Spectra
17	TRUE	2	40	2	1098	850	6
18	TRUE	2	15	21	1941	853	8
11	TRUE	2		22	1695	176	1
12	TRUE	2		7	2446	825	1
13	TRUE	3		22	2171	347	1
14	TRUE	3	35	19	2694	1313	0
20	TRUE	2	40	27	2271	826	2
21	TRUE	2	40	7	1978	1360	0
22	TRUE	2	100	5	2483	1501	0
1	TRUE	1	40	11	2315	1757	1
4	TRUE	1	15	3	1883	723	1
2	TRUE	1	85	9	2031	1329	0
9	TRUE	3	25	0	2286	792	1
6	TRUE	1	95	25	1300	500	0
5	TRUE	3	35	10	3039	1169	0
7	TRUE	1	55	18	2480	881	0
36	FALSE	2	90	8	1937	1666	5
37	FALSE	1	30	0	2587	2369	3
51	FALSE	3	40	0	2231	841	5
46	FALSE	3	35	7	2480	1276	4
40	FALSE	1	20	0	2414	130	1
47	FALSE	1	20	4	2144	1802	1
48	FALSE	1	50	13	2243	1955	6
49	FALSE	2	40	5	1477	27	4
56	FALSE	1	40	2	1178	609	0
57	FALSE	1	21	34	1857	1125	1
58	FALSE	1	32	5	2622	457	0
60	FALSE	1	24	9	1802	1077	1
61	FALSE	3	40	16	2270	1328	1
8	FALSE	3	5	2	3131	671	1
35	FALSE	1	75	14	2398	1922	1
32	FALSE	2			2160	1664	0
30	FALSE	2	30	9	1059	90	0
33	FALSE	2	60	15	2348	1630	2
65	FALSE	3	80	33	1361	1042	4
10	FALSE	2	30	5	2241	1136	0
29	FALSE	1	16	2	1875	207	0
52	FALSE	1	20	1	2162	1932	0
53	FALSE	1	40	2	2262	1356	0
54	FALSE	3	35	15	1411	1199	1
55	FALSE	2	30	2	1849	1245	2
44	FALSE	1	12	10	2283	760	1
38	FALSE	1	40	13	1900	1159	0

to perform better than others and it is of interest to validate that these trends can be reproduced in a different context. We will use the smaller vulvar cancer dataset to reproduce all the analyses included in Chapter 5 in order to investigate which trends are reproduced in this second dataset, and which are not. The results of classification on the vulvar cancer data are included in Appendix D, and discussed in Section 5.6.

The vulvar data, similarly to the endometrial cancer described above, consists of multiple sections from two TMAs, but fewer patients are represented in the vulvar TMAs as compared with the endometrial. The endometrial data represents information from 43 patients relevant to this study, but the vulvar data represents information from only 28 patients.

1.6 Statistics Background

As mentioned above, the data we will consider is of peaklist form. Peaklist data does not fit into the standard statistical paradigm of 'observations' and 'variables' in an obvious way, as the data simply consist of peaks with m/z values, several other characteristics (such as peak height or intensity, integrated area, and SNR), and labels indicating the spectra from which the peaks originated.

In each application we construct observations and variables from the peaklists in a different way appropriate to the particular context. In many contexts our observations correspond to spectra as is common in the analysis of MALDI-MSI data, but in the endometrial data for example our observations correspond to patients. The variables will always be some discretisation of the m/z domain — where each variable corresponds to an m/z interval and peaks are grouped based on which variable or m/z interval they fall into. We will present different approaches to discretisation of the m/z domain, each with advantages and disadvantages, and will use different approaches in different applications. The values of these variables for any given spectrum then correspond to a particular characteristic of the peaks in that spectrum within the relevant m/z interval. Characteristics that can correspond to the values of the variables include: the intensity of a peak, the integrated area of the peak, SNR of the peak, and binary values coding the presence or absence of a peak. In cases when multiple peaks from a single spectrum occur in the same m/z interval, these values can be averaged, or the maximum value taken, but we will choose sufficiently small m/z intervals such that multiple peaks from the same spectrum should not occur in the same m/z interval. In Section 2.3.1 we discuss how to choose m/z intervals that are sufficiently small such that multiple peaks from the same spectrum do not occur in the same m/z interval. We will consider each of these options for values of the variables in different applications.

In each of the options discussed above, the decisions about how to discretise the m/z domain into variables and what objects correspond to observations, allow us to represent the data as a $d \times n$ data matrix $\mathbb{X}$, where rows correspond to variables and columns to observations. Representing these data in this form allows us to draw on the methods from the multivariate statistics literature. Two such methods we will consider in some detail are clustering, and classification. We introduce these two methods, along with some associated concepts and a little discussion of the literature, in Section 1.6.1 and Section 1.6.2 respectively. Note that we will repeat much of the discussion from Section 1.6.2 in Section 4.1, where we discuss the particular classification methods we will use in more detail. Similarly we will repeat some of the discussion from Section 1.6.1 in Section 2.2.2, where we discuss

the particular clustering method we will use in more detail.

1.6.1 Clustering, Similarity, and Distance

Clustering is a field in which the objective is to group (usually partition) observations into 'clusters' such that observations in the same cluster are more 'similar' to each other than observations from different clusters. See Jain et al. (1999) for a review on clustering.

In different applications the term 'similar' can be interpreted in different ways but a general way to discuss it is in terms of a distance — two observations are 'similar' if the distance between them is small. Distance can be defined more precisely, for example as a pseudometric (Definition 1).

Definition 1. Distance (Pseudometric): *a non-negative function D such that for a set S, $D : S \times S \to \mathbb{R}$ and for every $x, y, z \in S$,*

1. $D(x, x) = 0$,
2. $D(x, y) = D(y, x)$, *and*
3. $D(x, z) \leq D(x, y) + D(y, z)$.

The word distance is sometimes used to mean any function $D : S \times S \to \mathbb{R}$ such that smaller values of $D(x, y)$ indicate x and y are more 'similar', for some intuitive interpretation of the term 'similar'. This more vague usage of the word 'distance' is more general and includes functions that are not pseudometrics but could still be meaningfully used in any of the places we use pseudometrics. That said, all the distances we use are pseudometrics, so in the context of this thesis the two usages of the word 'distance' are interchangeable and we simply use the more precise pseudometric definition (Definition 1) to avoid confusion.

Although we use several other distances later, in the context of clustering we focus on three distances in particular: the Euclidean distance (or L^2 norm), the cosine distance (Definition 2), and the Hamming distance (Definition 3).

Definition 2. Cosine Distance: $D_{cos} : \mathbb{R}^d \times \mathbb{R}^d \to [0, 2]$, *such that*

$$D_{cos}(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\sqrt{(\boldsymbol{x} \cdot \boldsymbol{x})(\boldsymbol{y} \cdot \boldsymbol{y})}}. \tag{1.3}$$

Definition 3. Hamming Distance: $D_{Ham} : \{0, 1\}^d \times \{0, 1\}^d \to [0, d]$, *such that*

$$D_{Ham}(\boldsymbol{x}, \boldsymbol{y}) = d - \boldsymbol{x} \cdot \boldsymbol{y} - (1 - \boldsymbol{x}) \cdot (1 - \boldsymbol{y}). \tag{1.4}$$

The Euclidean distance allows 'similar' to be interpreted as 'spatially nearby', and as such is useful in many clustering applications. Although there are more general definitions for the Hamming distance than Definition 3, we restrict the domain to d-dimensional binary vectors $\{0, 1\}^d$, as this is the context in which we will use the Hamming distance. It is interesting to note that, in this restricted domain of binary vectors, the Hamming distance is actually equivalent to the squared Euclidean distance, i.e. $D_{Ham} = D_{Euc}^2$.

There are many different approaches to clustering, see Koch (2013, Chapter 6) and references therein. One popular choice is to use Markov chain Monte Carlo methods to estimate mixture model parameters — for references to this approach see McLachlan and Basford (1988); Peel and McLachlan (2000). Xu et al. (2005) review more clustering approaches, including so-called 'fuzzy' clustering and neural network based methods. One approach to clustering of particular note due to its popularity in the analysis of MALDI-MSI is that described by Deininger et al. (2008), which combines 'supervised' and 'unsupervised' steps, and is described as as 'semi-supervised' by Deininger et al. (2008). Due to the popularity of such semi-supervised approaches in the MALDI-MSI context we will briefly describe the approach of Deininger et al. (2008), including a brief discussion of agglomerative hierarchical clustering that forms a component of it, although we do not pursue the idea further. Finally we will briefly discuss the k-means approach, which we discuss in more detail in Section 2.2.2, as this is the clustering approach we will primarily rely on in this work.

Agglomerative hierarchical clustering is an approach in which each observation can be thought of as beginning in a separate cluster, and an iterative process is implemented such that at each iteration two clusters, chosen due to being the 'most similar' in some sense, are combined. The total number of clusters is reduced by one in each iteration. This process ends either when a predefined number of clusters is reached or after $n - 1$ steps when all observations are in the same cluster. Hierarchical clustering such as this can be visualised in a dendrogram or hierarchical tree, describing the order in which clusters were combined. Sometimes the dendrogram representing a hierarchical clustering is informative, such as in the context described by Winderbaum et al. (2012). In the context of MALDI-MSI such a dendrogram is difficult to interpret due to the spatial information not being represented in the dendrogram. A specific partition of the observations into clusters can be plotted spatially and thus interpreted usefully. In order to produce a specific partition using standard hierarchical clustering methods a predefined number of clusters needs to be chosen, but justifying this choice can be difficult.

So-called 'semi-supervised' methods popular in the analysis of MALDI-MSI data, such as that described by Deininger et al. (2008), produce a particular set of clusters by first visualising a standard, unsupervised, agglomerative clustering in a dendrogram and then allowing the user to recursively open branches in the dendrogram — essentially performing a 'supervised' divisive clustering step where the choice of which cluster to split in each case is determined by the user. Beginning with the single cluster containing all observations the user 'opens' clusters, splitting them in two, and thus effectively moves down through the dendrogram (splitting the observations into more and more clusters) until the user believes they have reached a meaningful set of clusters. What the user determines to be a meaningful set of clusters is usually judged on the basis of the histology of the tissue with the aid of a H&E stain as described in Section 1.3.1, but this process is subjective nonetheless. This combination of unsupervised hierarchical clustering and user-determined steps is what earns these methods the term 'semi-supervised'. We will not consider results from such semi-supervised methods, or any hierarchical clustering methods at all, instead favouring the k-means approach described below, which compared to hierarchical methods is more computationally efficient.

k-means is an alternative approach to that of hierarchical clustering that attempts to find an optimum partition of the observations into k clusters such that observations in the same cluster are 'more similar' than observations in different clusters in some sense. We give details on the k-means algorithm in Section 2.2.2.

In comparison to the semi-supervised approaches described above, k-means is an entirely unsupervised method. We make extensive use of k-means to separate tissue types in the ovarian cancer data, where we know reasonable values of k to be the number of different tissue types present, which is easily determined from the H&E stains.

These different clustering approaches each have advantages and disadvantages, and both hierarchical and k-means clustering are used in the analysis of MALDI-MSI data, as Jones et al. (2012) discuss. We favour the k-means approach, as its unsupervised nature allows us to repeat many computations in parallel, and it is more computationally efficient than hierarchical methods as it does not need to generate full dendrograms. However it should be noted that hierarchical approaches have their uses as well, as discussed at length by Alexandrov (2012) and references therein.

1.6.2 Classification

Classification, sometimes called discriminant analysis, is a field that concerns the differentiation of data belonging to several classes (Koch, 2013, Chapter 4). Classification can be subdivided into two steps:

- Constructing a classification rule capable of assigning a class label to an observation. The construction of the rule is done on the basis of data with known class membership (labels, identifying the class to which each observation belongs), sometimes called 'training' data. This construction step is often referred to as the 'training' or 'learning' step.

- Applying a rule to assign a class label to an observation (or observations). This step can be further subdivided into one of two cases:

  - applying the rule to an observation of unknown class membership for which a real-world decision needs to be made (prediction), or

  - applying the rule to observations of known class membership in order to assess the performance of the rule (sometimes called 'testing').

It should be noted that there are a plethora of approaches to classification, and we only consider a very limited selection, which we describe in Section 4.1. Popular approaches to classification not represented in this work include random forests (see Breiman (2001) and references therein) and Support Vector Machines (SVMs) (see Schölkopf and Smola (1998); Cristianini and Shawe-Taylor (2000) and references therein). Special consideration of the analysis of binary data has dated back as far as Cox (1972), and although we consider classification of binary data, we do not consider classification approaches specifically developed for binary data, although such approaches do exist (see Lee and Jun (2011) and references therein). Comparison of a broader selection of classification approaches for binary data, such as that of Asparoukhov and Krzanowski (2001) is needed, but is beyond the scope of this work. Comparisons of a broader range of classification approaches (without restricting to binary data) have been made on MALDI-MS data by Wu et al. (2003) and on cDNA/mRNA microarray data by Dudoit et al. (2002) but an extensive comparison of classification approaches on MALDI-MSI data specifically is lacking. Extensive comparisons of classification approaches is lacking, but there has been significant interest in the classification of MALDI-MSI data in the literature (Casadonte and Caprioli, 2011). Some research of particular interest from the literature includes:

- Mascini et al. (2015) suggest using Principal Component Analysis (PCA)-Linear Discriminant Analysis (LDA) for the classification of MALDI-MSI data, and this is one of the options we consider in Chapter 5.

- Rauser et al. (2010) consider SVM and Artificial Neuronal Network approaches to the classification of top-down (protein level) MALDI-MSI data of fresh frozen tissue — a fundamentally different approach to the bottom up (peptide level) data from FFPE tissue that we consider.

- Casadonte and Caprioli (2011) discuss the importance of classification applied to MALDI-MSI data acquired on TMAs constructed from FFPE tissue. The endometrial cancer data we consider falls in this category, as do the data considered by Groseclose et al. (2008) and Djidja et al. (2010), who apply univariate dimension reduction approaches followed by SVM classification and PCA based classification respectively.

Although there is need for a comparison of a broader range of classification approaches for MALDI-MSI data, this is beyond the scope of this work — we aim to fulfil a similar role to the work of Djidja et al. (2010), Mascini et al. (2015) and Groseclose et al. (2008) on the classification of TMA MALDI-MSI data. Each of these papers considers the application of a particular classification approach on each of their respective datasets. We consider the application of a slightly wider variety of approaches, comparing the results of these approaches and suggesting approaches that seem to be promising on the basis of these results.

We will restrict attention to two-class problems and linear classification approaches, but it should be noted that multi-class and non-linear alternatives exist and are simply not discussed here. As we will only deal with linear two-class classification, we introduce some notation in Equation 1.5 specific to that scenario, and which we will use as a framework for comparing classification approaches within these restrictions.

First, let us consider the canonical Fisher's LDA for some intuition on linear classification. In the context of linear classification approaches, Fisher (1936) proposed to project the data onto a direction which leads to the best separation into two parts of the one-dimensional projected data. Equivalently Fishers proposal results in a vector of weights, $\boldsymbol{d}$, such that the linear combination of the data with this vector yields one-dimensional quantities which, ideally, fall into two disjoint intervals. In practice the projected one-dimensional data may not separate completely into separate intervals, and an offset or scalar value β is used to represent the value that achieves the best separation.

To state this more formally, first let $\mathbb{X}$ denote a $d \times n$ data matrix of n observations with known class labels coded as -1 or $+1$. All the rules we will consider use the data $\mathbb{X}$ and the associated class labels to 'train' a rule by finding a $d \times 1$ vector $\boldsymbol{d}$ and a scalar β. This rule then assigns class label $\tau(\boldsymbol{x})$ to a $d \times 1$ observation $\boldsymbol{x}$ in the following way:

$$\tau(\boldsymbol{x}) = \begin{cases} +1 & \text{if} \quad \boldsymbol{d}^T\boldsymbol{x} + \beta > 0 \\ -1 & \text{if} \quad \boldsymbol{d}^T\boldsymbol{x} + \beta < 0 \end{cases} . \tag{1.5}$$

The notation of Equation 1.5 is repeated in Equation 4.1, where we explore these ideas in more detail and also present an example application of Fisher's LDA to illustrate the intuition behind these concepts. The difference between the classification approaches we discuss in Sections 4.1.2, 4.1.3 and 4.1.4 essentially boils down to different approaches to choosing $\boldsymbol{d}$ and β.

Chapter 2

DIPPS and Exploratory Analyses

MALDI-MSI can produce large amounts of data (Bonnel et al., 2011). It also has a uniquely complicated structure. The development of novel statistical tools is required in order to analyse and interpret MALDI-MSI results as many standard methods are inappropriate. Deininger et al. (2008), Bonnel et al. (2011), Alexandrov et al. (2013), and references therein have suggested a number of approaches to the exploratory analyses of MALDI-MSI data. Here we suggest an approach based on feature selection from the binary data. Our approach is based entirely on the peaklist data. Dealing only with the peaklist data, as opposed to the full spectrum data, provides a reduction in the quantity of information involved in computations by several orders of magnitude. Our approach provides easily interpretable results, fast computation, and we suggest it for use as an 'initial pass' for quality control as well as a starting point for more in-depth follow-up analyses. The approach we suggest, and its application discussed in Section 3.1, has been published as Winderbaum et al. (2015). Here, we take the opportunity to discuss the ideas in more detail.

This chapter introduces and illustrates methods for:

- Representing MALDI-MSI datasets as data matrices in a way that is robust and useful, facilitating further analyses.
- Separating meaningful subsets in the data by the use of clustering methods.
- Characterising subsets of the data in an easily interpretable way using the DIPPS.

Addressing the first of these dot points, binning as a method for representing MALDI-MSI data in binary binned form is reviewed in Section 2.1, accompanied by details and related concepts in Appendix A. k-means clustering is reviewed in Section 2.2 as an exploratory method that can be used to investigate structure in these data. The k-means clustering approach we discuss can be seen as an alternative to the so-called 'semi-supervised' methods of Deininger et al. (2008); Alexandrov et al. (2010) and Bonnel et al. (2011) who cluster MALDI-MSI data using principal component analysis, hierarchical clustering, and Gaussian mixture models. In Section 2.3 the binning and k-means methods are illustrated and practical concerns related to their use are discussed in detail. In Section 2.4 I introduce the DIPPS, and how it can be used to extend and further investigate the results of k-means clustering of the binary data. We use the DIPPS in a feature selection approach conceptually similar to that of Jones et al. (2011), but taking a very different approach in practice, focusing on the binary data in a way that allows for the result to be visualised in a single, easy to interpret, image. In Section 2.5 I introduce a novel spatial smoothing algorithm for binary data that can aid in the analysis and visualisation of binary

data and for sparse data can even act as a dimension reduction step. In Section 2.6 I illustrate the use of the DIPPS, as introduced in Section 2.4, for investigating the ovarian cancer data and demonstrate how the DIPPS is of practical use due to its intuitive and powerfully simple interpretation.

The methods in this chapter do not only apply to MALDI-MSI data but can be used more generally on any binary presence/ absence data as introduced in Definition 8. In order to clarify the generality of these methods, I introduce methods in a general sense and then motivate them with examples from the MALDI-MSI application in separate sections. For example, I introduce the k-means method generally in Section 2.2.2 and provide motivating applications of k-means to MALDI-MSI in Section 2.3.3. Similarly I introduce the general concepts underlying the DIPPS and related ideas in Section 2.4, motivating these ideas with examples in the MALDI-MSI context in Section 2.6.

Throughout this chapter I will use the dataset A3 as a motivating example of the ovarian cancer datasets of Section 1.5.1 when illustrating the application of methods to MALDI-MSI data. All analyses illustrated on dataset A3 in this chapter have been duplicated on all the ovarian cancer datasets, but results will often only be shown for the motivating dataset A3 because the focus of this chapter is on single-dataset analyses. Comparisons between multiple datasets is the focus in Section 3.1, where the results for the remainder of the ovarian cancer datasets are summarised and discussed.

2.1 Binning

In this section I discuss binning as it applies to peaklist MALDI-MSI data. As discussed in Section 1.6, peaklist data does not fit into the standard statistical paradigm of 'variables' and 'observations. Binning, in the context that I will use it, is a method for constructing variables from peaklist data. As mentioned in Section 1.6, this construction of variables is achieved by discretising the m/z range into intervals and grouping peaks whose m/z values fall in the same interval. Binning corresponds to a particular choice of discretisation — equal width intervals that partition the m/z range. We choose the bin locations arbitrarily, i.e. data-independently, and as such any dataset we analyse will have the same bins. The data-independent nature of this binning allows for single-dataset analyses to be extended to multiple-dataset comparisons in a natural way as the datasets will have the same variables, the variables having been constructed from the same bins.

Because of its data-independent nature, the binning we suggest (Algorithm A.1) has the inherent disadvantage of sometimes placing bins in the 'wrong' place. Alternative, data-dependent, methods can improve on this by using the data to inform decisions on where to place bins. We introduce a data-dependent discretisation method and mention some alternatives in Section 3.2, and apply this approach to the glycan data. However, the extension of such data-dependent methods from single to multiple dataset analyses is not as intuitive and can be more computationally intensive than it is for binning. Comparison of multiple datasets is the main interest in the ovarian cancer data introduced in Section 1.5.1, and is often important in the context of MALDI-MSI data more broadly, and so we favour the data-independent binning approach for these data.

Binning is widely used, but for completeness and to avoid ambiguity I include explicit definitions in Appendix A, as well as some extended notation that will be of use in the discussion to follow. One such discussion point is how to make

sensible choices about the size of bins, which we consider in Section 2.3.1 through the exploration of the motivating dataset A3.

2.2 k-means Clustering

Often it is of interest to separate subsets of a dataset that are 'different' from each other in some sense. For example in a MALDI-MSI dataset there could be a number of different tissue types represented. Mass spectra collected from the same tissue type should, in principle, be relatively similar while spectra from different tissue types should be relatively less similar. A natural first step in the analysis of such data is to verify that this intuitive statement is supported by the data — i.e. to pose the questions: "Is it possible to separate groups of similar spectra?" and if it is, "How do these groups compare with the different tissue types we expect to see based on the histology?". The similarity, or conversely dissimilarity/ distance, between two spectra is fundamentally important in addressing such questions. We discussed some of the ideas relating to measuring similarity/ distance in Section 1.6.1, and we will continue to use these ideas here. There are many approaches to clustering, as discussed in Section 1.6, and here we review the k-means approach. k-means is an iterative method that attempts to find a partition of the observations into groups such that the variability within groups is as small as possible. One of the keys to measuring such within-group variability is that a representative 'centroid' vector can be found for any set of vectors, and then the 'variability' of that set can be measured by the sum of distances from individual vectors to their centroid.

This section is organised as follows. First I introduce the definition of a centroid, and its form in the context of some common distances. Secondly I will introduce the k-means clustering algorithm explicitly, and discuss some of the decisions that must be made in implementing it. The application of k-means to the ovarian cancer data will be discussed in Section 2.3.

2.2.1 Centroids

As mentioned above, k-means attempts to find a partition of the observations into groups such that the variability within groups is as small as possible. The concept of a distance (Definition 1) gives us a way to quantify the dissimilarity of two vectors. One way of quantifying the variability in a group of observations is to find a 'centroid' vector, representative of that group, and sum the distances from each observation in the group to the centroid vector.

Definition 4. Centroid: *Given a distance D and a $d \times n$ matrix $\mathbb{X}$ with columns denoted $\boldsymbol{x}_{\bullet j}$, the centroid, $\boldsymbol{x}^*$, of the $\boldsymbol{x}_{\bullet j}$ is*

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \mathbb{R}^d}{\arg\min} \left\{ \sum_{j=1}^{n} D(\boldsymbol{x}, \boldsymbol{x}_{\bullet j}) \right\}.$$

The interpretation of a centroid will vary depending on the distance used, for example for a set of real-valued vectors:

- When using the Euclidean distance the centroid is the component-wise arithmetic mean of the vectors.

- When using the cosine distance the centroid is the mean of the vectors after having been normalised to be equal length.
- When using the Hamming distance, the centroid is the component-wise median of the vectors. Note that we limit the domain of the Hamming distance to binary vectors, see Definition 3.

A centroid need not be unique. The centroid when using the Euclidean distance is unique, and when the number of vectors in the set is odd, the centroid for the Hamming distance is unique. The centroid when using the cosine distance however is only unique up to multiplication by a constant. As mentioned above, the centroid can be used to quantify the variability in a set of vectors as the sum of distances, or squared distances, from the centroid. It is interesting to note that such a sum is unique, even when the centroid is not.

2.2.2 k-means Algorithm

k-means, as described in Koch (2013, Section 6.3), is a method for partitioning n vectors into k 'clusters' such that the vectors in each cluster are similar by some distance. Here we discuss the implementation of k-means clustering,

Algorithm 2.1. **k-means:** *Given a $d \times n$ matrix $\mathbb{X}$ with jth column denoted $\boldsymbol{x}_{\bullet j}$ a distance D, and k initial cluster seeds $\boldsymbol{x}^*_{1[0]}, \boldsymbol{x}^*_{2[0]}, \ldots, \boldsymbol{x}^*_{k[0]}$, perform the following steps at the s^{th} iteration:*

1. *Calculate $c_j = \arg\min_{\kappa} \left\{ D(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}^*_{\kappa[s-1]}) \right\}$ for $j = 1, 2, \ldots, n$.*
2. *Calculate the centroids $\boldsymbol{x}^*_{\kappa[s]}$ of $\{\boldsymbol{x}_{\bullet j} \,|\, c_j = \kappa\}\, \forall \kappa = 1, \ldots, k$, see Definition 4 in Section 2.2.1.*

Stop when $\boldsymbol{x}^*_{\kappa[s]} = \boldsymbol{x}^*_{\kappa[s-1]}\, \forall \kappa \in \{1, 2, \ldots, k\}$. The c_j of the last iteration denote the resulting cluster membership of the observations $\boldsymbol{x}_{\bullet j}$.

Algorithm 2.1 can be sensitive to the choice of the initial cluster centroids. For different choices of initial seeds, the algorithm can converge to different solutions. This is due to the algorithm getting "stuck" in local minima. There are two approaches to addressing this:

1. Use Algorithm 2.1 many times with different, random, initial cluster seeds and use the solution with minimum sum of to-centroid distances.
2. Fix the initial cluster seeds. If the initial cluster seeds can be justified, this is often an attractive option due to the deterministic reproducibility of results.

Note that when using Euclidean distance in the context of k-means we have actually used the squared Euclidean distance, but these are actually equivalent for many purposes. The sum of to-centroid distances as in point 1. above is one of the only things this choice actually affects in a meaningful way.

2.3 Preliminary Analysis of Dataset A3

In this section I present an exploratory analysis of dataset A3, applying binning and k-means as reviewed in Section 2.1 and Section 2.2 respectively. In Section 2.3.1 I

discuss the application of binning in practice, in particular choice of bin size. In order to interpret the results of k-means clustering, spatial patterns should be considered. Thus, in Section 2.3.2 I introduce how to represent the results of clustering spatially. Finally, in Section 2.3.3 I consider the results of k-means clustering on dataset A3. I use the k-means method as both an exploratory technique and also as a means to assess different options for proceeding with the analysis of these data. Specifically, I use the k-means results presented in Section 2.3.3 to discuss two major choices that I will carry through all analyses that follow for the ovarian cancer datasets:

- The data type to use: binary data, or non-binary data such as intensity, area or SNR. I conclude that using the binary data is appropriate for the analysis of the ovarian cancer datasets, although I revisit this choice in Section 4.2 for the endometrial cancer datasets.

- The distance to use in the k-means clustering: I consider Euclidean, Hamming, and cosine distances as options and conclude that the cosine distance produces the most stable results overall.

2.3.1 Choice of Bin Size

The first step in our approach to the analysis of MALDI-MSI peaklist data is the discretisation of the m/z domain. This discretisation allows for distinct variables to be constructed from the continuous m/z range by grouping peaks that are nearby in m/z. Here we will use binning — specifically Algorithm A.1 — for this discretisation. Algorithm A.1 requires a bin size parameter b to be chosen, which specifies the width of each m/z region, or bin, in Dalton (Da). Here we explore different choices for the bin size b, ultimately reaching the conclusion that $b = 0.25$ is a reasonable choice for these peptide MALDI-MSI data. Analyses that follow on from this section use this bin size of 0.25.

Choosing an appropriate bin size is a balance between two competing objectives:

1. Peaks originating from different molecular species should be placed in different bins. The smaller the bin size, the more likely we are to achieve this objective.

2. Peaks originating from the same molecular species should be placed in the same bin. The larger the bin size, the more likely we are to achieve this objective.

The same molecular species should not have more than one peak per spectrum, so we can assume that two peaks from the same spectrum should originate from different molecular species. We can use the minimum distances between peaks within the same spectrum to estimate the minimum distance between peaks from different molecular species. If we then select a bin size smaller than the minimum distance between peaks originating from different molecular species, we are guaranteed to achieve the first objective. We will then choose a bin size as large as possible in order to maximise the chance of achieving the second objective as well.

In order to consider the distances between peaks within spectra more closely, let us denote the set of these intra-spectrum differences

$$\mathbb{D} = \left\{ m_{(i)j} - m_{(i-1)j} \,|\, i \in [2, N_j],\, j \in [1, n] \right\}, \tag{2.1}$$

where N_j denotes the number of peaks in spectrum j and $m_{(i)j}$ denotes the m/z of the ith peak in spectrum j, where the peaks are sorted in increasing order of m/z

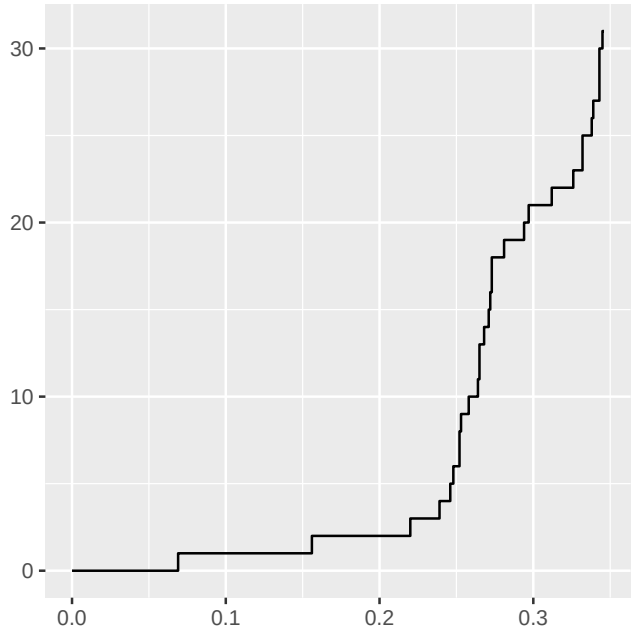


Figure 2.1: Plot of d_t on the y-axis, as defined in Equation 2.2, against t on the x-axis for dataset A3.

within each spectrum. Let d_t be the number of intra-spectrum differences $m \in \mathbb{D}$ below a threshold t, so

$$d_t = |\{m \in \mathbb{D} | m < t\}|, \tag{2.2}$$

given the notation for $\mathbb{D}$ in Equation 2.1. Note that d_t in Equation 2.2 denotes the cardinality or size of the set $\{m \in \mathbb{D} | m < t\}$. Figure 2.1 shows d_t for small values of t in the dataset A3.

Although these data are of very high quality, allowing for a small rate of false positive peaks is still reasonable. In this context, a false positive peak would be a peak resulting from instrument or chemical noise — not originating from a molecular species in the sample — or appearing at an incorrect m/z. If we consider that $\mathbb{D}$ contains 1287670 differences, we see that $d_{0.25} = 6$ is a comparatively tiny number, and so it is reasonable to say that using a bin size of 0.25 (almost) achieves the first objective of peaks from different molecular species always occurring in different bins. Consideration of Figure 2.1 leads to the impression that there is an 'elbow' in the cumulative number of differences (d_t) at approximately 0.25, meaning that if a bin size is increased much beyond 0.25, the number of intra-spectrum peak-pairs that will be in the same bin will rapidly increase. Given these two heuristic arguments, we conclude that using a bin size of 0.25 is a reasonable compromise between the two competing objectives. Despite this justification for the choice of bin size, all analyses that follow in this chapter have been replicated with a range of bin sizes, and these results are quite robust to changing the choice of bin size within the range $0.05 - 3$.

It is interesting to note how this heuristic approach relates to the discussion of Section A.4:

- Definition 18 corresponds to no two peaks from the same spectrum being placed in the same bin.

- Definition 19, which guarantees Definition 18 for bin sizes less than b^*, in this

context is

$$b^* = \arg\max_t \{d_t | d_t = 0\} = 0.069,$$

i.e. the maximum t such that $d_t = 0$, as seen in Figure 2.1.

2.3.2 Visualising spatial distribution of an imaging dataset

As discussed in Section 1.4, MALDI-MSI is the process of acquiring spectra from many spatially distributed points across the surface of a tissue sample. In order to interpret results we will want to represent the results spatially. Representing results spatially will allow us to compare results with histological features; for example finding differences between tumour tissues and healthy tissues as these will be spatially separated and visible in the histology. Figure 2.2 shows the spatial distribution of the positions from which spectra have been acquired in dataset A3. In dataset A3, and in all the MALDI-MSI datasets we consider, spectra were collected from a regular grid across the tissue. In large droplet analyses, such as that considered by Gustafsson et al. (2015), it is common for positions to be unevenly distributed across the tissue, in order to represent the tissue types being targeted. However, in the high lateral resolution spray approach to MALDI-MSI that we consider, the whole tissue section is systematically mapped and so the locations from which spectra are acquired are systematically distributed in a regular grid. I will use the spatial distribution of Figure 2.2 to visualise results in the following sections. For example, the result of clustering is a cluster membership for each spectra in the dataset. These types of results can be visualised spatially by plotting the pixels shown in Figure 2.2, and colouring them according to the cluster membership — pixels of the same colour belonging to the same cluster, pixels of different colours belonging to different clusters. Such a plot can then easily be compared with stained tissue images, and relationships between the cluster membership and tissue morphology inferred, as in Figure 2.6.

In Section 2.3.1 I briefly mentioned the number of adjacent peak pairs within spectra in dataset A3, $d_\infty = 1287670$. d_t was defined in Equation 2.2. This number can be seen to originate from the fact that dataset A3 consists of 1301720 peaks detected over 14050 spectra, as noted in Table 1.2, and $d_\infty = 1301720 - 14050$. In actual fact, spectra were collected from 14059 locations, but no peaks were detected in 9 of these spectra, also noted in Table 1.2. In the context of cluster membership results these 9 empty spectra mentioned above would be greyed out as they are not assigned a cluster membership. If you look very closely you will notice 9 interior grey pixels in Figure 2.2 — these correspond to the 9 so-called 'empty' spectra in which no peaks were detected.

2.3.3 Results of k-means clustering

In this section I will present results of k-means clustering, as described in Section 2.2.2, for dataset A3. I will use these k-means clustering results to compare and discuss the choice of which form of the data, binary or various non-binary forms, and which distance to use. In order to conduct these k-means clusterings a number of other choices must be made, specifically:

- Bin size,
- Number of clusters,
- Initial cluster seeds, and

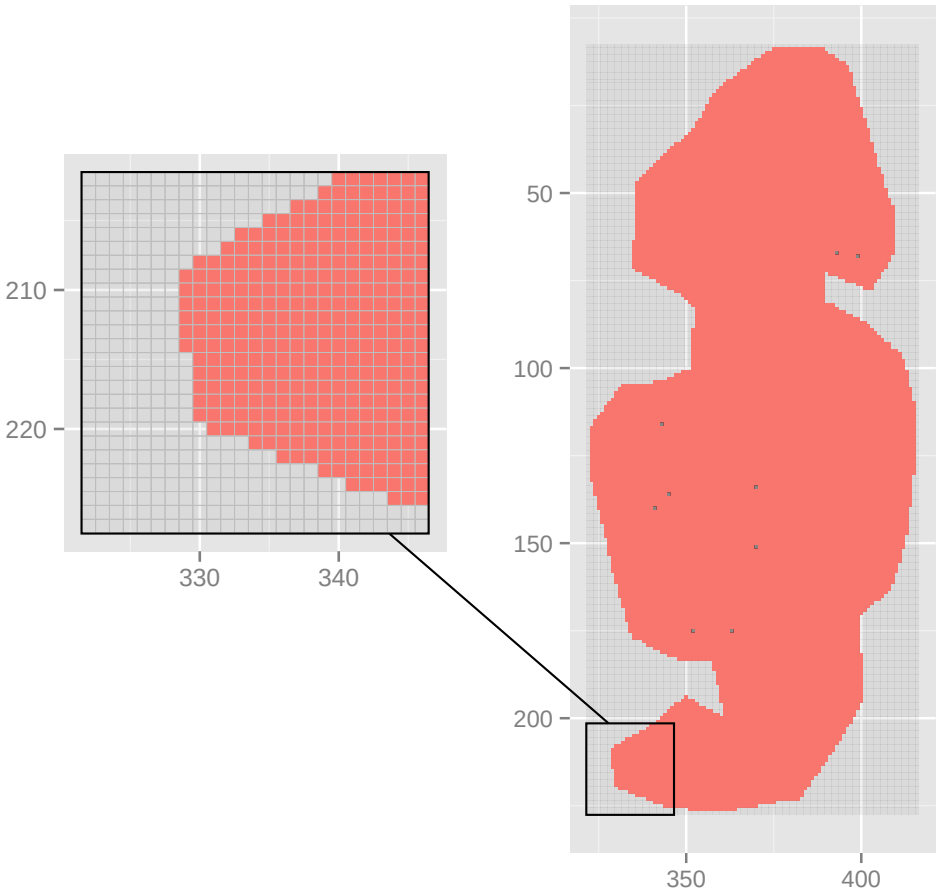


Figure 2.2: Spatial distribution of spectra in dataset A3. Coloured pixels indicate X-Y coordinates from which spectra were collected. The zoom-in highlights that the coloured region consists of square pixels arranged in a regular grid, each pixel at X-Y coordinates where a spectrum was collected.

- How to deal with formation of empty clusters.

and so first we will provide a brief discussion for our choices regarding these four points.

In Section 2.3.1 I discussed the reasoning for choosing a bin size of 0.25, and I continue with this choice of bin size here and in all following analyses. Nevertheless, we repeated all analyses with a variety of bin sizes in the range 0.05–3 in order to verify robustness to small changes in bin size, and saw no noticeable effect. It has previously been shown that four is an appropriate number of clusters for these data — see Koch (2013, Example 6.12, Section 6.5.3). Based on the histology, as shown in the H&E stain of Figure 2.6, it is expected that three broadly different tissue types should be observed in this dataset, corresponding to cancer, adipose and stroma. Off-tissue spectra makes for four broadly different spatial regions. I will present results here using number of clusters $k = 4$. Analyses where the value of k is varied have also been considered, but are omitted for brevity. In the interest of reproducibility I choose initial cluster seeds from the observations in a deterministic, automated manner that selects observations that achieve extrema when projected into the first few principal component directions. PCA is described in more detail in Section 4.3.1. Similarly to the other choices, the robustness of results was established by considering clustering on the basis of different seeds — for example, the clustering results shown in Winderbaum et al. (2015) are generated by performing 100 k-means clusterings in parallel each with initial cluster seeds chosen from the observations at random, and using the clustering that resulted in the lowest sum of to-centroid distances of the 100 resulting clusterings. Sometimes all observations will be allocated to a strict subset of clusters in step 1. of Algorithm 2.1, causing $\{\boldsymbol{x}_{\bullet j} \,|\, c_j = \kappa\}$ to be empty for some κ. This is a problem, as it causes Algorithm 2.1 to fail at the following step 2. We solve this problem by introducing a 'singleton' cluster of a single observation corresponding to the observation with the greatest to-centroid distance whenever an empty cluster is formed. This solution also guarantees that Algorithm 2.1 will always produce a cluster membership with exactly k clusters, and is implemented in the MATLAB `kmeans` function.

As discussed in Section 2.3.2 I will represent the results of k-means clustering as spatial maps of the cluster-membership using colours to distinguish clusters — each spectrum represented as a coloured pixel at its X-Y coordinates. The results of k-means clustering on the binary representation of the data, using three different distances, are shown in Figure 2.3. The results of k-means clustering on three different types of non-binary data are shown in Figure 2.4 using the Euclidean distance, and in Figure 2.5 using the cosine distance.

We expect cluster memberships to be spatially localised due to the nature by which the data was collected — adjacent spectra are collected from adjacent areas of tissue, and are expected to be more similar than spectra from arbitrary locations. It can be seen in Figure 2.3 that the clusters produced using the binary data are quite spatially localised. The clusters produced using the Euclidean and cosine on the binary data (Figure 2.3) agree well and seem to, despite a small amount of "speckling", match up with the morphology of the tissue as shown by the H&E stain in Figure 2.6, colours roughly corresponding to: cancer, stroma, adipose, and off-tissue. Interpretation of the clusters produced by the Hamming distance is less clear, combining most of the cancer regions with the off-tissue. The Hamming distance could be detecting the cancer tissue as being similar to the off-tissue regions due to the spectra acquisition being poor on cancer areas of tissue and less signals being detected — causing there to be similarity between the two by their shared

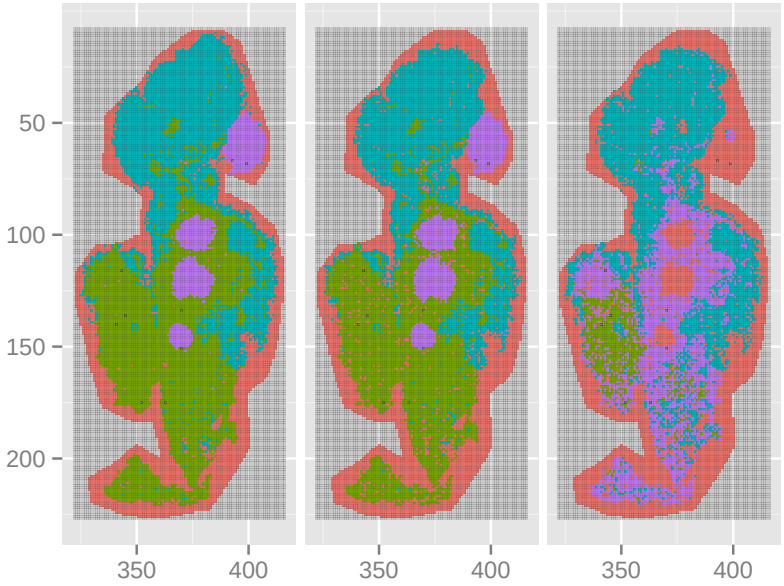


Figure 2.3: The results of 4-means clustering (Algorithm 2.1) on the dataset A3. Analysis was done using the cosine (left), Euclidean (centre) and Hamming (right) distances on the binned (bin size 0.25) binary data. Apart from minor speckling effects, spatially localised clusters that well separate the main tissue types are apparent in all three clustering results.

absence of many signals. We pursue the discussion of the cancer regions being characterised by the absence of certain signals further in Section 2.6. Although the Euclidean and cosine clusterings appear to be similar based on this visual inspection of Figure 2.3, comparisons across the remainder of the ovarian cancer datasets show that the cosine distance clustering tends to pick out regions corresponding to tissue types more consistently, and I will continue to use the cosine distance in further analyses. The results of applying the cosine clustering to the remainder of the ovarian cancer datasets are discussed in Section 3.1.

We are interested in comparing the results obtained from binary and non-binary data. Figure 2.4 shows the results of k-means clustering by Euclidean distance on three variants of non-binary: intensity, area, and SNR data. By considering Figure 2.4, it is immediately clear that the clusters produced using the Euclidean distance on the non-binary data fail to separate the tissue types known to be present in the tissue. This failure strongly contrasts the results on the binary data shown in Figure 2.3. At first glance this may seem worrying, as the non-binary data contain strictly more information than the binary representation. The fact that the non-binary data fail to separate tissue types could be a cause of serious concern about data quality. On further consideration however, the fact that the binary data contains a strict subset of the information in the non-binary data, and successfully separates tissue types, seems to indicate that the additional information contained in the non-binary data adds 'noise' that obscures the information capable of separating tissue types. Methods for the removal of this unwanted 'noise' could be considered, such as that discussed in Section 4.4 for the endometrial cancer data. For the ovarian cancer data however, using the binary representation of these data bypasses the issue of noisy measurements, and we will consider the binary data in further analyses. The results of k-means clustering using the cosine distance on the

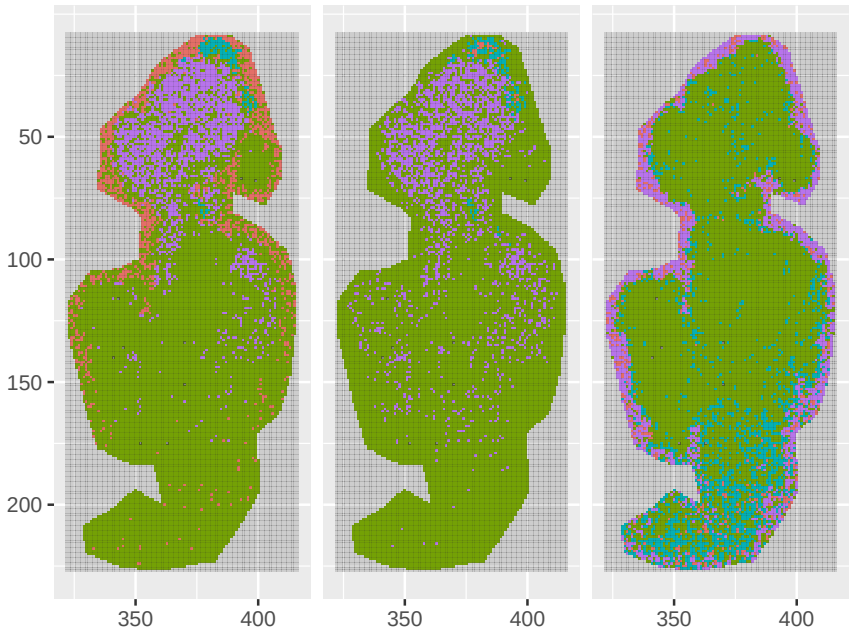


Figure 2.4: The results of 4-means clustering (Algorithm 2.1) on the dataset A3. Analysis was done using the Euclidean distance on binned (bin size 0.25); intensity (left), area (centre) and SNR (right) data. All three clustering results show spatially de-localised clusters. Although the clustering results show various degrees of success at separating certain features of the tissue, all fail to separate the main tissue types present in the tissue section.

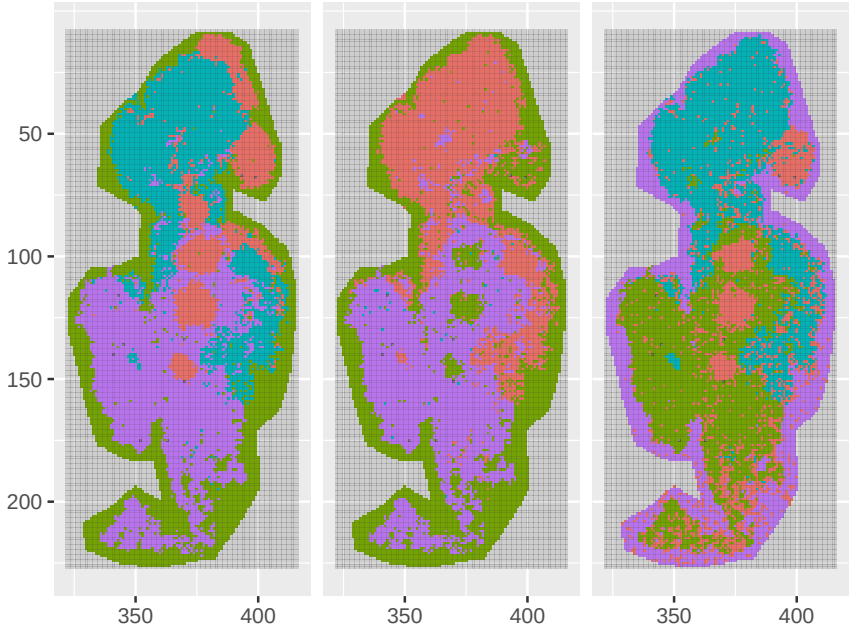


Figure 2.5: The results of 4-means clustering (Algorithm 2.1) on the dataset A3. Analysis was done using the cosine distance on binned (bin size 0.25); intensity (left), area (centre) and SNR (right) data.

same three types of non-binary data are shown in Figure 2.5. The cosine distance results of Figure 2.5 show much improvement over their Euclidean distance counterparts shown in Figure 2.4, able to separate some of the tissue types. The non-binary cosine distance results of Figure 2.5 still separate tissue types worse than the binary data results of Figure 2.3, despite the cosine distance showing much improved performance in comparison to the Euclidean distance. The performance of the cosine distance in Figure 2.5 contributes to our decision to use the cosine distance, as well as the binary data, in further analyses.

2.4 Feature Extraction for Binary Data

Often we know that a given subset of data is particularly interesting or has some meaningful interpretation. For example, as in Section 2.3, we may discover that a cluster seems to correspond to cancerous tissue, and want to further investigate it. It can be desirable to identify variables that distinguish or characterise such a subset. In this section we introduce our approach to identifying such variables, as published in Winderbaum et al. (2015), specifically:

- The Difference in Proportions of Occurrence Statistic (DIPPS).
- A heuristic cut-off value for the DIPPS that allows variable ranking by DIPPS to be used to automate feature extraction of meaningful subsets of variables.

In Section 2.6 we will demonstrate and further motivate these concepts through their application to dataset A3, including how easily interpretable and concise heatmap visualisations of such selected subsets of variables can be constructed and used for the interpretation of MALDI-MSI data. First I introduce some notation for subsets of data in Section 2.4.1. Then I introduce the concepts relating to the DIPPS as summarised above in Section 2.4.2.

2.4.1 Subset Notation

Given that rows and columns of matrices are usually assigned a unique index between 1 and n, where n is the number of rows or columns respectively, using an n-index subset (Definition 5) to denote subsets of rows or columns is convenient.

Definition 5. n-Index Subset: *A set $\mathcal{C}$ is an n-index subset if and only if $\mathcal{C} \subset \{1, 2, \ldots, n\}$.*

A useful duality exists between n-index subsets and binary vectors of length n, see Definition 6. This dual notation will be useful for writing many subset operations in a concise matrix form.

Definition 6. Binary Vector Dual (of an n-Index Subset): *Let $\mathcal{C}$ be an n-index subset as in Definition 5. $\mathbf{c}$ is the binary vector dual of $\mathcal{C}$ if $\mathbf{c}$ is a length n binary vector with the value 1 at every location whose index is contained in $\mathcal{C}$ and value 0 elsewhere.*

Binary vectors are prominent in this work, and it is interesting to note that operations on these binary objects could also be formulated in terms of their set dual and boolean algebras, but this idea is not explored here as it does not contribute to the objective of our work.

Let the n-index subset $\mathcal{C}$ correspond to a subset of the columns of a data matrix $\mathbb{X}$. There exists a transformation matrix such that when post-multiplied by $\mathbb{X}$, the resulting matrix is the corresponding sub-matrix of $\mathbb{X}$, see Definition 7.

Definition 7. Subset Transformation Matrix: *Let $\mathcal{C}$ be a n-index subset as in Definition 5. Let $\mathfrak{c}$ be the binary vector dual of $\mathcal{C}$ as in Definition 6. Let the number of ones contained in $\mathfrak{c}$ be denoted $n_{\mathcal{C}}$, and let the $n_{\mathcal{C}}$ non-zero entries of $\mathfrak{c}$ occur at indices i_k for $k = 1, 2, \ldots, n_{\mathcal{C}}$.*

The subset transformation matrix $T_{\mathfrak{c}}$ for the n-index subset $\mathcal{C}$ is the $n \times n_{\mathcal{C}}$ binary matrix such that the k^{th} column of $T_{\mathfrak{c}}$ contains the value 1 in the i_k^{th} position and zeros elsewhere.

Notice that

$$\mathfrak{c} = T_{\mathfrak{c}} 1_{n_{\mathcal{C}} \times 1}.$$

We use this n-index subset notation to represent sub-matrices.

2.4.2 DIPPS

In this section we introduce DIPPS and related concepts in a general context. We motivate these ideas briefly and in a general setting in order to emphasise that the DIPPS approach is general and could be applied to any data of a presence/ absence binary type. Often MALDI-MSI observations will correspond to spectra, variables to m/z bins. In the particular application we will use to illustrate these ideas — the ovarian cancer data of Section 1.5.1 — the subset of interest will usually be the cancer tissue, its complement being surrounding non-tumour healthy tissues and off-tissue spectra. These ideas and how they can be interpreted in these specific cases are explored in more detail in Section 2.6, but in this section we aim to introduce the concepts in general without this specific context.

First we define a couple of terms.

Definition 8. Presence/ absence data *is any binary data whose two values, numerically coded one and zero, can be interpreted as the presence or absence of some characteristic.*

Binary binned MALDI-MSI data as produced by Algorithm A.1 is presence/ absence as per Definition 8 — the two values coding for the presence/ absence of some characteristic (a peak) in a given variable (m/z bin) and observation (spectrum).

A natural question to ask of presence/ absence data is 'how many observations demonstrate presence of a particular characteristic?'. Proportions of occurrence are a natural way to measure this in a way such that allows sets of different sizes to be comparable.

Definition 9. Proportion of Occurrence *The mean of a set of presence/ absence observations can be interpreted as the proportions of the observations in which the characteristic is present.*

In the same way that proportions of occurrence measure how many observations exhibit presence of a characteristic, DIPPS measures the degree to which a subset of observations differs from its complement in the proportion of occurrence of some characteristic. Let us consider a $d \times n$ presence/ absence data matrix $\mathbb{X}$ in which we are interested in finding variables that distinguish between a particular subset of observations, $\mathbb{X}T_{\mathfrak{c}}$, and the rest of the data, $\mathbb{X}T_{(\mathbf{1}_{d\times 1} - \mathfrak{c})}$. DIPPS is the difference between the proportions of occurrence in these two subsets of the data.

Definition 10. Difference in Proportions of Occurrence Statistic (DIPPS) *Let $\mathbb{X}$ be a $d \times n$ binary data matrix. Let $\mathcal{C}$ be an n-index subset with binary vector*

dual $\mathfrak{c}$ *and size* $n_{\mathcal{C}}$*. The DIPPS for a row of* $\mathbb{X}$ *is the corresponding element of the* $d \times 1$ *vector*

$$\boldsymbol{\rho}(\mathcal{C}) = \left(\frac{1}{n_{\mathcal{C}}}\mathbb{X}\mathfrak{c}\right) - \left(\frac{1}{n - n_{\mathcal{C}}}\mathbb{X}\left(\mathbf{1}_{d\times 1} - \mathfrak{c}\right)\right)$$

The first term of $\boldsymbol{\rho}$ in Definition 10, $\frac{1}{n_{\mathcal{C}}}\mathbb{X}\mathfrak{c}$, is the vector of proportions of occurrence within the subset of interest, $\mathbb{X}T_{\mathfrak{c}}$. The second term, $\frac{1}{n-n_{\mathcal{C}}}\mathbb{X}\left(\mathbf{1}_{d\times 1} - \mathfrak{c}\right)$, is the vector of proportions of occurrence for its complement, $\mathbb{X}T_{(\mathbf{1}_{d\times 1}-\mathfrak{c})}$. If you think of the 'presence' value as predicting membership in the subset of interest of an observation, then these two terms can be thought of as measures of sensitivity and specificity respectively. From this prediction perspective, each of the d DIPPS in Definition 10 is a combined measure of both sensitivity and specificity for the corresponding variable. This can be thought of as a cost function for which false positives and false negatives are weighted equally, and this special equally-weighted case is sometimes called the informedness — see Powers (2011); Fawcett (2006) and references therein. Variables, or rows of $\mathbb{X}$, with a higher proportion of occurrence in the subset of interest than in its complement will have positive DIPPS. Variables with a higher proportion of occurrence outside of the subset of interest than in it will have negative DIPPS. These variables can be referred to as *positive indicators* (where the 'presence' value predicts membership) and *negative indicators* (where the 'absence' value predicts membership) for the subset of interest $\mathcal{C}$ respectively. A variable that has the same proportion of occurrence in the subset of interest as outside of it will have a DIPPS of zero.

The DIPPS of Definition 10 provides an intuitive ranking of variables by their ability to characterise/ predict a given subset of observations $\mathcal{C}$. Choosing a cut-off value allows us to select a subset of variables using this ranking by selecting variables with DIPPS above the cut-off. We suggest a heuristic for choosing an appropriate cut-off in Definition 12, but first we need some notation for which variables are selected using a given cutoff, and for that we use the concept of a DIPPS-template.

Definition 11. DIPPS-Template: *Let* $\mathcal{C}$ *be an* n*-index subset, and* $\boldsymbol{\rho}$ *be the corresponding vector of DIPPS as in Definition 10. For a given cut-off value* a*, the positive (negative) DIPPS-template* $\boldsymbol{t}_{a^+}$ *(*$\boldsymbol{t}_{a^-}$*) is a* $d \times 1$ *binary vector, each element of which is* 1 *if the corresponding element of* $\boldsymbol{\rho}$ *is* $\geq a$ *(*$\leq -a$*) and* 0 *otherwise.*

Note that although $\boldsymbol{\rho}$, $\boldsymbol{t}_{a^-}$, and $\boldsymbol{t}_{a^+}$ of Definition 11 are functions of $\mathcal{C}$, I omit this dependence as in this context $\mathcal{C}$ is assumed to be fixed. Without loss of generality I will discuss positive DIPPS-templates. Given a cutoff value a for the DIPPS ranking of the variables in some $d \times n$ data, the positive DIPPS-template is the $d \times 1$ vector of indicator variables for the DIPPS of the corresponding variables being above the cutoff a. The variables selected are those whose proportion of occurrence in $\mathcal{C}$ are least a greater than the proportion of occurrence in the complement of $\mathcal{C}$. One way to interpret this is that a randomly chosen observation from $\mathcal{C}$ is 'at least a more likely' to exhibit presence than a randomly chosen observation from the complement of $\mathcal{C}$. Extending this thinking, the DIPPS-template can be thought of as a 'representative' observation from $\mathcal{C}$ — exhibiting presence in variables for which the DIPPS is at least a. Recalling the concept of a centroid from Section 2.2.1, which is a general approach to constructing a 'representative' vector for a set of vectors, but not restricted to being binary in general. The idea behind the heuristic DIPPS-threshold is that we choose a such that these two approaches to constructing representative vectors are

the most similar — i.e. so that the DIPPS-threshold is as similar to the centroid of $\mathcal{C}$ as possible.

Definition 12. DIPPS-Threshold: *Given: a $d \times n$ binary data matrix $\mathbb{X}$; an n-index subset $\mathcal{C}$ (with binary vector dual $\mathfrak{c}$); and a distance D (Definition 1), let $\boldsymbol{c}$ denote the centroid (Definition 4) of the subset of interest ($\mathbb{X}T_{\mathfrak{c}}$), then for positive (negative) indicators, the DIPPS-threshold a_*^+ (a_*^-) is defined as:*

$$\textit{For positive indicators}: \qquad a_*^+ = \arg\min_a \left\{ D\big(\boldsymbol{c}, \boldsymbol{t}_{a^+}\big) \right\}$$

$$\textit{For negative indicators}: \qquad a_*^- = \arg\min_a \left\{ D\big(\boldsymbol{c}, \boldsymbol{1}_{d\times 1} - \boldsymbol{t}_{a^-}\big) \right\}$$

The DIPPS-threshold of Definition 12 provides a natural way of obtaining a set of positive indicators for $\mathcal{C}$ (variables with DIPPS $\geq a_*^+$) and a set of negative indicators for $\mathcal{C}$ (variables with DIPPS $\leq -a_*^-$). These sets of variables are identified by the entries of $\boldsymbol{t}_{a_*^+}$ and $\boldsymbol{t}_{a_*^-}$ equal to one respectively. These sets are useful as they quickly and easily provide a shortlist of variables that distinguish the subset of interest and can be investigated further in follow-up analyses. We explore the use of DIPPS in generating shortlists of variables for follow-up analyses in Section 2.6. In Section 3.1 we consider another use of DIPPS — comparing the sets of indicators generated from different datasets in order to separate within-patient from between-patient variability.

The DIPPS-threshold of Definition 12 has interesting properties when particular distances are used, but in general attempts to maximise the similarity between the DIPPS-template of Definition 11 and the centroid of the subset $\mathcal{C}$ of the data. If we begin with the empty set the corresponding DIPPS-template would be $\boldsymbol{t}_{\infty^+} = \boldsymbol{0}_{d\times 1}$. Usually adding the variable with the highest DIPPS, and then the second highest, and so on decreases the distance between the DIPPS-template and the centroid $\boldsymbol{c}$. Eventually adding more variables will begin increasing this distance, and the DIPPS-threshold of Definition 12 attempts to find the cutoff, and corresponding DIPPS-template/ subset of variables that achieves the local optimum for which the template to centroid distance is minimised.

In order to clarify the concept of the DIPPS-threshold of Definition 12, I will briefly consider how it applies when the cosine distance is used. The cosine distance is the distance we focus on following from the discussion of Section 2.3.3, and so this choice will be particularly relevant. I also limit attention to the positive indicator case a_*^+, as this will be the one we focus on and as the positive indicator case is equivalent to the negative indicator case up to swapping ones/ zeros or presence/ absence.

When the cosine distance is used (i.e. $D = D_{cos}$),

$$D_{cos}\big(\boldsymbol{c}, \boldsymbol{t}_{a^+}\big) = 1 - \frac{\boldsymbol{t}_{a^+}^T \boldsymbol{c}}{||\boldsymbol{c}||.||\boldsymbol{t}_{a^+}||}$$

Let n_a denote the number of non-zero entries of $\boldsymbol{t}_{a^+}$ and let $b_2 < b_1$ such that $n_{b_2} = n_{b_1} + 1$. As the cutoff a becomes smaller, changing from b_1 to b_2, $D_{cos}\big(\boldsymbol{c}, \boldsymbol{t}_{a^+}\big)$ will decrease if $(\boldsymbol{t}_{b_2^+} - \boldsymbol{t}_{b_1^+})^T \boldsymbol{c} < \left(\boldsymbol{t}_{b_1^+}^T \boldsymbol{c}\right)\left(\sqrt{\frac{n_{b_2}}{n_{b_1}}} - 1\right)$ and will increase if $(\boldsymbol{t}_{b_2^+} - \boldsymbol{t}_{b_1^+})^T \boldsymbol{c} > \left(\boldsymbol{t}_{b_1^+}^T \boldsymbol{c}\right)\left(\sqrt{\frac{n_{b_2}}{n_{b_1}}} - 1\right)$. The left-hand side of these expressions is the entry of the

centroid $\boldsymbol{c}$ corresponding to the variable added when changing the cutoff from b_1 to b_2. The right-hand side is a function of the total number of variables selected so far, n_{b_1}. As n_{b_1} increases, $\left(\sqrt{\frac{n_{b_2}}{n_{b_1}}} - 1\right)$ approaches zero and when this term becomes sufficiently small reducing the cutoff a further only increases the distance $D_{cos}\left(\boldsymbol{c}, \boldsymbol{t}_{a^+}\right)$. This limiting behaviour is what allows for a local minima to be found.

2.5 Spatial Smoothing for Binary Data

All of the methods introduced so far completely ignore the spatial information present in MALDI-MSI datasets. One way to incorporate such distance meta-data is through a spatial smooth, however when smoothing binary data it is desirable to maintain the binary form of the data through the smoothing process. Many common smoothing algorithms, such as most kernel and polynomial spline smooths for example, do not preserve the binary nature of the data. In this section I propose a spatial smooth (Algorithm 2.2) that preserves the binary nature of the data through use of cellular automata. For a broader perspective and history on cellular automata see Mitchell et al. (1996); Wolfram (1984) and references therein. The approach I propose to smoothing is to our knowledge novel, and has now been published in Winderbaum et al. (2015).

We represent the spatial information associated with a $d \times n$ binary data matrix $\mathbb{X}$ as a $n \times n$ distance meta-data matrix $\mathfrak{D}$, whose rows and columns correspond to the same observations represented by the columns of $\mathbb{X}$. A distance meta-data matrix $\mathfrak{D}$ is a symmetric matrix such that the $(i,j)^{th}$ entry of $\mathfrak{D}$ is the value of a distance D between the i^{th} and j^{th} observations in the dataset, i.e. $D(\boldsymbol{x}_{\bullet i}, \boldsymbol{x}_{\bullet j})$. This implies that the diagonal of $\mathfrak{D}$ is filled with zeroes, as $D(\boldsymbol{x}_{\bullet i}, \boldsymbol{x}_{\bullet i}) = 0 \; \forall i$. In MALDI-MSI data, we will choose the distance to be the Euclidean distance between the X-Y coordinates of spectra, so the physical distance between the spatial locations of two spectra. We use the lateral resolution, i.e. the minimum distance between two spectra or the width of a pixel in spatial maps such as in Section 2.3.2, as the unit of measurement. The smooth we propose in Algorithm 2.2 looks at a spatial neighbourhood around each observation or spectrum and if enough of the neighbouring observations differ the value is changed to agree with the neighbourhood. This process is then repeated iteratively until a stable state is found in which every neighbourhood meets the minimum agreement criteria. This iterative process thus guarantees the resulting data will meet a minimum level of 'smoothness'.

Algorithm 2.2. Spatial Smooth: *Given: a smoothing parameter $0 \leq \tau < \frac{1}{2}$; a distance cutoff $\delta > 0$; a stopping point $\tilde{k}$; a binary $d \times n$ data matrix $\mathbb{X}$, and a $n \times n$ distance meta-data matrix $\mathfrak{D}$. Initially let $\mathbb{X}^{(0)} = \mathbb{X}$. For $k = 1, 2, \ldots$ construct $\mathbb{X}^{(k)}$ by the following steps:*

1. *For all j, find the $1 \times n$ binary vector $\mathfrak{c}_j$ such that each element of $\mathfrak{c}_j$ is one if the corresponding element of the j^{th} row of $\mathfrak{D}$ is $\leq \delta$ and zero otherwise.*

2. $$x_{ij}^{(k)} = \begin{cases} x_{ij}^{(k-1)} & \text{if} \quad \left(1 - x_{ij}^{(k-1)} + \left(2x_{ij}^{(k-1)} - 1\right) \frac{\boldsymbol{x}_{i\bullet}^{(k-1)} \mathfrak{c}_j^T - x_{ij}^{(k-1)}}{\mathbf{1}_{1\times n} \mathfrak{c}_j^T - 1}\right) > \tau \\ 1 - x_{ij}^{(k-1)} & \text{if} \quad \left(1 - x_{ij}^{(k-1)} + \left(2x_{ij}^{(k-1)} - 1\right) \frac{\boldsymbol{x}_{i\bullet}^{(k-1)} \mathfrak{c}_j^T - x_{ij}^{(k-1)}}{\mathbf{1}_{1\times n} \mathfrak{c}_j^T - 1}\right) \leq \tau \end{cases}$$

Stop when either $k = \tilde{k}$ or $\mathbb{X}^{(k)} = \mathbb{X}^{(k-1)}$. When one of the stopping conditions is reached, $\mathbb{X}^{(k)}$ is the spatially smoothed data.

Remarks on Algorithm 2.2:

- The term $\left(1 - x_{ij}^{(k-1)} + \left(2x_{ij}^{(k-1)} - 1\right) \frac{\boldsymbol{x}_{i\bullet}^{(k-1)} \boldsymbol{c}_j^T - x_{ij}^{(k-1)}}{\mathbf{1}_{1\times n} \boldsymbol{c}_j^T - 1}\right)$ in step 2. is the proportion of observations in a δ-neighbourhood of the j^{th} observation $\boldsymbol{x}_{\bullet j}^{(k-1)}$ that have the same value as $\boldsymbol{x}_{\bullet j}^{(k-1)}$ for their i^{th} variable. If there is an insufficient proportion of neighbouring similar observations (specifically $< \tau$), the value $x_{ij}^{(k-1)}$ is 'smoothed' to agree with the majority of its neighbours, i.e. $x_{ij}^{(k)} = 1 - x_{ij}^{(k-1)}$. Otherwise, $x_{ij}^{(k-1)}$ remains unchanged at the kth step, i.e. $x_{ij}^{(k)} = x_{ij}^{(k-1)}$.

- We choose $\delta = \sqrt{2}$ which results in a range 1 *Moore neighbourhood*. See Gray (2003). This neighbourhood is used in the cellular automata literature including Conway (1970). It is worth noting that acquiring the range 1 Moore neighbourhood by using the Euclidean distance and $\delta = \sqrt{2}$ is equivalent to using the Tchebychev distance (L^∞ or sup-norm), and $\delta = 1$ so long as the observations lie on a regular grid. It is important to deal with the special case of empty δ-neighbourhoods, in which case we leave the data unmodified. However it is always possible to select δ sufficiently large such that there are no empty δ-neighbourhoods, and in our case $\delta = \sqrt{2}$ is sufficiently large to satisfy this condition. Alternative neighbourhoods could be selected by choosing different combinations of cutoff δ and distance D, but we have not explored these possibilities.

- The smoothing parameter τ defines the proportion of neighbouring spectra needed to agree in order for the value of an observation to remain unchanged at any given step, as discussed in the first point above. Small values of τ smooth less ($\tau = 0$ leaves the data unmodified), while larger values smooth more. The limit $\tau \to \frac{1}{2}$ results in maximum smoothing, and is equivalent to the intuitive median smooth. The median smooth tends to yield over-smoothed data in the case of MALDI-MSI data, and often fails to converge. We choose an intermediate smoothing parameter, $\tau = \frac{1}{4}$, for these analyses. For data on a regular grid results will not significantly change if τ is within the same $\frac{1}{8}$-wide interval, as changing τ within these intervals will affect only spectra on the boundary of the acquisition region (spectra with less than 8 neighbours). The values $\frac{1}{8}$ and $\frac{3}{8}$ could also be used, but in Section 2.6 and Section 3.1 we present results using the intermediate value $\tau = \frac{1}{4}$.

- Alternative smoothing options include kernel methods (Wand and Jones, 1995) which apply to the more general class of continuous data. These methods typically produce continuous values when applied to binary data, for which there is no clear interpretation. Our method produces binary smoothed data — maintaining the interpretability of the binary values. As one of the main strengths of using the binary data is its simple interpretation as 'presence/ absence' data, the ability to preserve this interpretation is important.

- At each smoothing iteration k, variables are smoothed independently, and within each variable all observations are smoothed simultaneously at each step. This means that it is possible to parallelise the smoothing algorithm, making relatively efficient use of computational resources.

- In practice the stopping point $\tilde{k}$ is not usually necessary, as typically convergence is reached in < 20 iterations. However it is good practice to include $\tilde{k}$

in case convergence is not reached, as we cannot easily guarantee convergence. An alternative use for $\tilde{k}$ is to improve computation speed — by choosing a small $\tilde{k}$, such as 2 or 3 for example, computation could be performed very quickly.

2.6 Characterisation of Cancer in Dataset A3

Here we illustrate the methods and ideas introduced in Sections 2.4 and 2.5 by applying these ideas to the motivating dataset A3 from the ovarian cancer data of Section 1.5.1. I will demonstrate the usefulness of these methods in the exploration of MALDI-MSI data, focussing on the interpretation of results. More specifically I will show how DIPPS can be used to find variables (m/z bins) that are important in distinguishing clusters, how such information can be presented as heatmaps and how these heatmaps visualise clustering results in a way that is easier to interpret than simply plotting the cluster membership.

Continuing on from the analyses of dataset A3 in Section 2.3, I will consider the 4-means clustering by cosine distance of the binary binned data, shown in Figure 2.6 side-by-side with an image of the same section of tissue after H&E staining. As mentioned in Section 2.3, a notable feature of Figure 2.6 is that through comparison with the histological staining, experts concluded that the clusters roughly correspond to the different tissue types present, specifically: off-tissue, adipose, cancer, and stroma. In Section 2.6.1 I begin with some discussion and exploratory analysis of how presence/ absence data and proportions of occurrence, as in Definition 8 and Definition 9 respectively, apply in this context. I will then consider the extraction of variables by use of DIPPS in Section 2.6.2. Finally in Section 2.6.3 I will show how heatmaps representing these extracted variables can provide a representation for the clustering results that allows for intuitive and meaningful conclusions to be easily drawn.

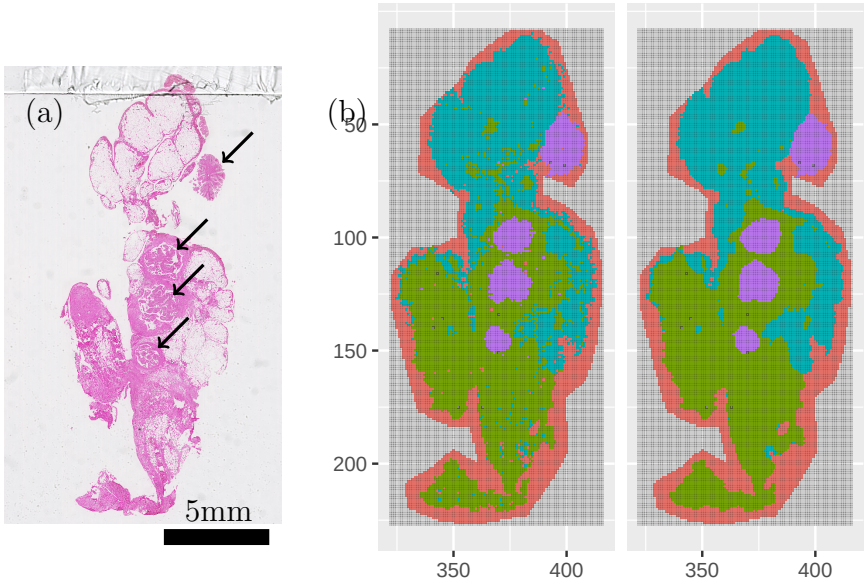


Figure 2.6: (a) H&E stained tissue section with arrows indicating the four visible tumours and (b) cluster membership resulting from 4-means clustering (Algorithm 2.1) by cosine distance on the binary binned (bin size 0.25) A3 data (Algorithm A.1) with smoothing ($\tau = 0.25$, right) and without smoothing ($\tau = 0$, left). Note that the left image in (b) reproduces the left-most image in Figure 2.3.

2.6.1 Proportions of Occurrence

In this section we present exploratory analyses of dataset A3 in order to develop familiarity with properties of these data so that we might interpret further results in appropriate context.

I begin by considering the distribution of proportions of occurrence (Definition 9) in the form of a histogram, as shown in Figure 2.7. The proportion of occurrence for each of the 4294 variables of the binned data were calculated, and Figure 2.7 shows a histogram of these proportions for the 4294 variables. Figure 2.7 shows a heavily right skewed distribution, with a large number of variables having a very low proportion of occurrence. This type of heavily right skewed distribution is typical for MALDI-MSI datasets such as this. To be explicit about the degree of right skew, 3271 or $\sim 76\%$ of 4294 variables have a proportion of occurrence less than 0.005 or 0.5% — that is to say that $\sim 76\%$ of non-empty bins contain peaks in less than 71 of 14050 spectra. A natural question to ask is:

"What happens to the clustering results if I remove these low-occurrence bins?"

If we construct a new data matrix $\mathbb{X}^*$ from $\mathbb{X}$ by removing the rows of $\mathbb{X}$ corresponding to bins with proportion of occurrence < 0.005 and perform a 4-means clustering on the modified matrix (using the same starting points) the results are almost indistinguishable — differing from the clustering results on the full unsmoothed data shown in Figure 2.6 by only 10, or 0.07%, of 14050 pixels. A continuous range of thresholds for cutting off the proportions of occurrence was considered but this illustrative example is sufficient to demonstrate the point that most of the low proportions of occurrence bins can be removed without significantly affecting the clustering.

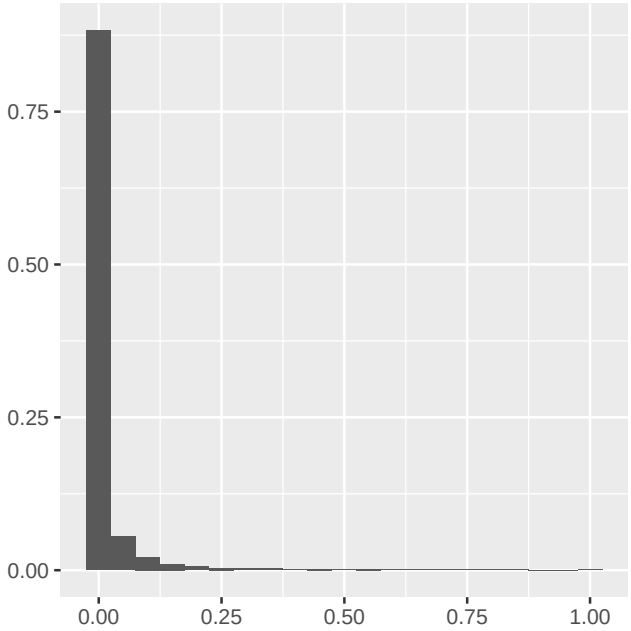


Figure 2.7: Histogram of the proportions of occurrence of the bins in the binary binned (bin size 0.25) A3 data. Proportion of occurrence is represented on the x-axis, relative frequency (relative to the total of 4294 bins) on the y-axis. The distribution shown is fairly typical for this type of data – heavily right skewed.

Low proportion of occurrence variables seem not to be important in successfully distinguishing tissue types by clustering. A large number of variables have a low proportion of occurrence, and this suggests that there exists a small subset of variables that can effectively distinguish tissue types. In order to understand what distinguishes tissue types in these data and identify potential subsets of variables that can do so, let us begin by considering the high proportion of occurrence variables. Figure 2.8 shows the spatial distribution of occurrence for each of the 8 highest proportion of occurrence variables. The m/z values that these images correspond to and their proportions of occurrence are included in the figure caption. Similarly to the representation of cluster-membership, I use the spatial distribution discussed in Section 2.3.2 to visualise the occurrence in a variable by using colour to distinguish the binary presence / absence values in the variable of interest.

As mentioned in Section 1.4 four internal calibrants (Gustafsson et al., 2012) were sprayed onto the tissue during sample preparation, for mass-calibration purposes. These calibrants should appear at known m/z values equally across the entire tissue section. Figure 2.8 (a), (b), (d), and (i) correspond to these calibrants. Figure 2.8 (c) and probably (g) are similarly trypsin autolysis products resulting from the trypsin sprayed evenly onto the tissue during sample preparation. Notice how the last of these calibrants, Figure 2.8 (i), exhibits a systematic absence for high Y-coordinate values (that is in the lower part of the image) and thus is not amongst the very highest proportion of occurrence m/z bins. Such a systematic absence indicates a problem, and the detection of such problems is part of the reason these internal calibrants are included — in order to further improve the methods used to acquire these data. This particular problem has as yet eluded explanation. This absence could potentially be an example of what I will refer to as 'false negative' measurements. These 'false negatives' could be caused by insufficient matrix being deposited, resulting in insufficient ions being produced and no peaks being detected. Peaks being mistaken for noise by the peak-picking algorithm could also account for false negatives. Essentially any situation where the molecule of interest is present in the sample, but not detected as a peak in the mass spectrum.

Figure 2.8 (f), (h), and (j) are interesting when considering the differentiation of tissue types, as their spatial distributions, visually, seem to match that of the cancer cluster in Figure 2.6. The visual match between the occurrence in the bins of Figure 2.8 (f), (h), and (j), and the cluster membership of Figure 2.6 is that of a negative indication for the cancer cluster — that is the *absence* of peaks in those bins seems to correspond to the cancer cluster. Positive indicators of cancer are far more useful than negative indicators, for two reasons:

- As briefly mentioned above, there is the issue of what I refer to as 'false negative' measurements, which make negative indicators somewhat dubious. 'False positive' measurements however are very rare, due to the fundamental nature of the data as described in Chapter 1, and so positive indicators are much more reliable.

- More importantly however, positive indicators for cancer are more useful from a *biochemical* perspective, as they have more potential as possible diagnostic and predictive tools, as tests for positive indicators can be devised with relative ease. Developing tests for the *absence* of something is somewhat more complicated and difficult, and the results would typically be less useful or informative.

So although these three negative indicators for the cancer are interesting, and could

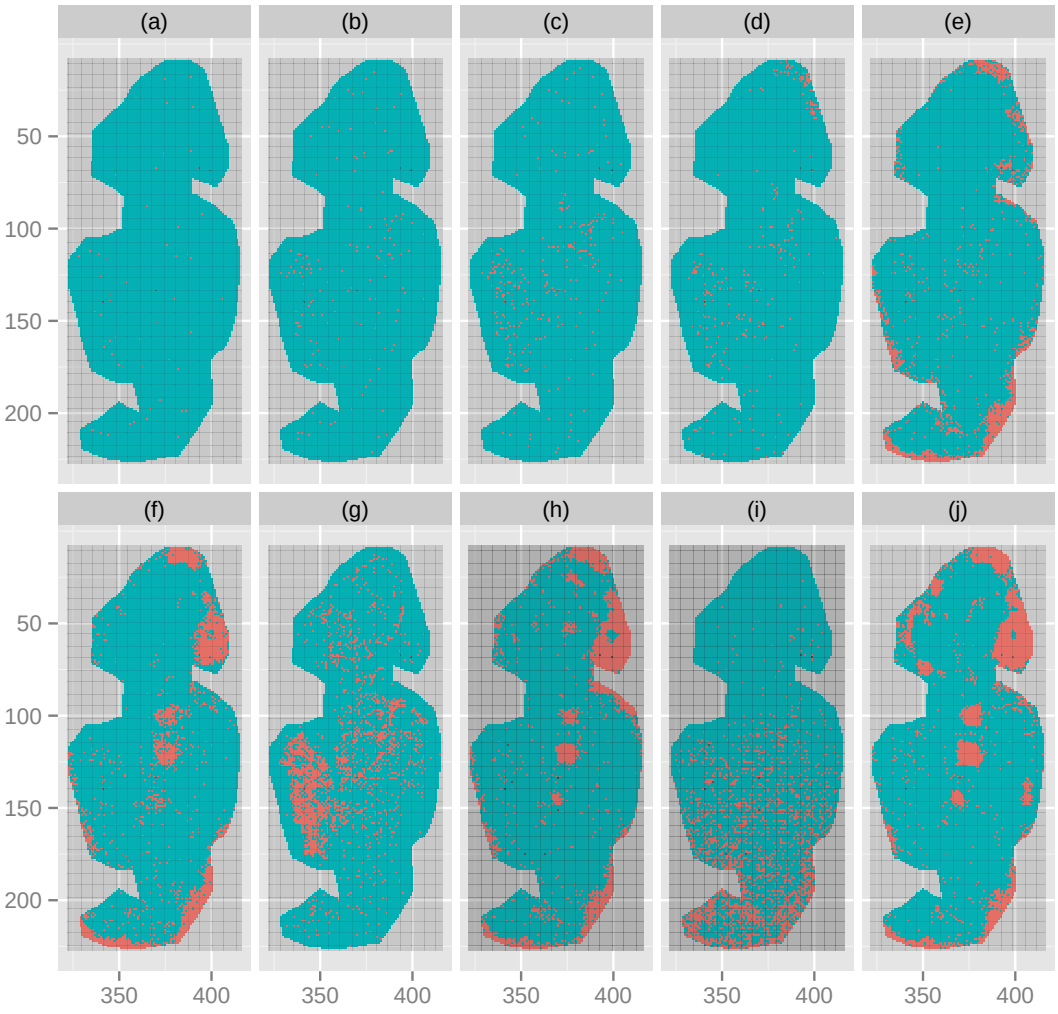


Figure 2.8: Occurrence maps for high occurrence m/z bins in dataset A3. Spatial maps indicate the presence / absence of peaks in each pixel (spectrum). Shown are the 0.25 Da wide bins centred at m/z:

(a) 2932.5 with occurrence 0.997,

(b) 2147.25 with occurrence 0.99,

(c) 2211 with occurrence 0.98,

(d) 1570.75 with occurrence 0.979,

(e) 1459.75 with occurrence 0.904,

(f) 1562.75 with occurrence 0.863,

(g) 2283.25 with occurrence 0.857,

(h) 1585.75 with occurrence 0.841,

(i) 1296.75 with occurrence 0.841,

(j) 1655.75 with occurrence 0.828.

be pursued by matching to LC data for identification and follow-up studies, we prioritise our efforts on pursuing positive indicators.

2.6.2 DIPPS-based Feature Extraction in Dataset A3

In Section 2.6.1 we mention that it is possible to distinguish tissue types on the basis of a small subset (less than a quarter) of the variables. We also discuss how a binary variable that characterises a cluster can do so in one of two ways: as a positive indicator, or a negative indicator. I will consider positive indicators, but all the methods introduced are general, and with minor modifications could be applied to negative indicators, or all indicators. In this section I demonstrate how the ideas introduced in Section 2.4 can be applied to find a subset of variables that characterise a cluster as positive indicators. I will show how such a subset of variables can be visualised in a useful way, and how this allows for a more interpretable representation of the clustering results than just the clustering results themselves.

The DIPPS of Definition 10 takes values between -1 and 1. If variables are listed in decreasing order of DIPPS, this ordering will rank variables from the best positive indicator to the worst positive indicator, or best negative indicator. Variables in the middle of this ranking, with DIPPS near zero, do not correlate with the cluster of interest. If we choose the cluster of interest to be the cancer cluster of the unsmoothed clustering shown in Figure 2.6, then Figure 2.9 shows variables with high DIPPS, i.e. good positive indicators for the cancer cluster — presence correlating with the cancer regions. Note that we could apply this DIPPS-based approach, and produce all the same results that follow from it in Section 2.6.3 to any of the clusters of Figure 2.6. We choose the cancer cluster from the unsmoothed result because we are particularly interested in the cancer, and this serves as an illustrative example. In Section 3.1 we consider the application of this approach in a more systematic way to all clusters, and we will limit our attention to the smoothed results. In general, considering spatial plots of all the variables in a MALDI-MSI dataset is tedious, and even for a shortlist such as this considering each image in detail is time consuming and does not yield meaningful interpretations. We present a method for combining the images of Figure 2.9 into a single heatmap in Section 2.6.3, and once follow-up analyses have been carried out on individual proteins then it is possible to come back and consider the individual spatial distributions of peptides originating from proteins known to be of interest. In order to obtain a subset of variables from this ranking, we use a threshold and take all variables with DIPPS above this threshold. We suggest a heuristic threshold, a_*^+, in Definition 12. For the unsmoothed cancer cluster of Figure 2.6 this heuristic $a_*^+ = 0.1832$. Figure 2.10 shows the dependence of $D(\boldsymbol{c}, \boldsymbol{t}_{a^+})$ on a, and the local minima at a_*^+. In dataset A3, 54 variables have a DIPPS greater than or equal to $a_*^+ = 0.1832$.

As mentioned in Section 2.1 and discussed further in Section 2.3.1 there are two main conditions we attempt to satisfy when we bin our data with Algorithm A.1:

- Peaks originating from different molecular species are placed in different bins, and
- Peaks originating from the same molecular species are placed in the same bin.

Realistically, due to the data-independent nature of Algorithm A.1, the second of these two goals can never be guaranteed to be completely satisfied. Either of these goals not being met can result in molecular species that would otherwise be detected as good positive indicators not being detected. We can address this issue by

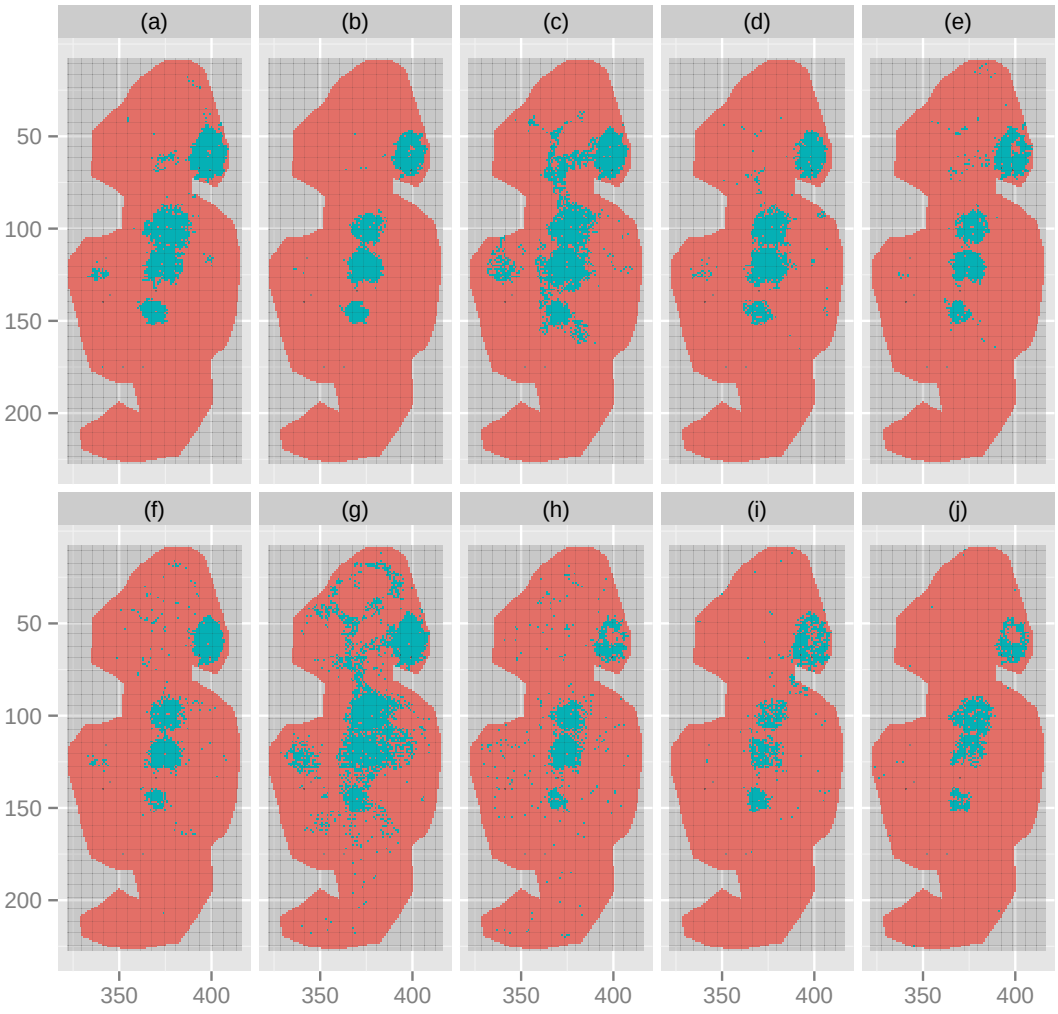


Figure 2.9: Occurrence maps for high difference in occurrence m/z bins in dataset A3. Spatial maps indicate the presence / absence of peaks in each pixel (spectrum). Shown are the 0.25 Da wide bins centred at m/z:

(a) 1406.75 with DIPPS 0.918,

(b) 2484.25 with DIPPS 0.883,

(c) 2854.5 with DIPPS 0.867,

(d) 1998 with DIPPS 0.859,

(e) 1936 with DIPPS 0.854,

(f) 1609.75 with DIPPS 0.848,

(g) 1390.75 with DIPPS 0.823,

(h) 2392.25 with DIPPS 0.717,

(i) 1740 with DIPPS 0.7,

(j) 2246.25 with DIPPS 0.657.

repeating the same analysis in parallel, but using Algorithm A.3 to shift the bin locations by half a bin-width. This ensures that each molecular species that ought to be detected will be detected in at least one of the two parallel analyses. We did this for dataset A3, and there are a number of things to note about the results of these shifted-bin analyses:

- The clustering is quite robust to the shifting bin locations. The unsmoothed clustering result of Figure 2.6 changes in only 277, or $\sim 2\%$, of 14050 pixels when repeated on the shifted-bin data.
- In the shifted-bin analysis, $a_*^+ = 0.1668$ for the cancer cluster, not very different to the $a_*^+ = 0.1832$ in the initial analysis.
- 54 variables (m/z bins) have a DIPPS greater than or equal to a_*^+ in the shifted-bin analysis — the same number as in the initial analysis. Of these 54 variables, 47 match between the two analyses in a one-to-one manner. Two pairs of adjacent bins in the initial analyses appear as a single bin in the shifted-bin analysis, and one pair of adjacent bins in the shifted-bin analysis appears as a single bin in the initial analysis. Three variables appear in the shifted-bin analysis that did not appear in the initial analysis, and two variables appear in the initial analysis that do not appear in the shifted-bin analysis. This behaviour highlights the importance of using multiple binnings in parallel in order to ensure important variables are not missed. In the interests of illustrating methods and ideas simply I will continue consider only one binning for the remainder of the discussion of the ovarian cancer data in this chapter and the analyses that follow in Section 3.1. I will revisit the concept of using multiple shifted-bin analyses in parallel in Section 4.2.1 when we consider classification of the endometrial data.

2.6.3 Visualising Characterisations of the Cancer Cluster

In Section 2.6.2 I demonstrated how the DIPPS can be used with a heuristic cutoff (Definition 12) to obtain a set of positive indicators for the unsmoothed cancer cluster of Figure 2.6. We will call these variables 'DIPPS-features', and in this section we will explore how the spatial distribution of these DIPPS-features can be visualised in an easily interpretable, and therefore useful, way.

We count how many DIPPS-features exhibit presence in each spectrum — that is, if we represent the DIPPS-features as a d-index subset of the variables with binary vector dual $\mathfrak{d} = \boldsymbol{t}_{a_*^+}$, then we consider the sum $\mathfrak{d}^T\mathbb{X}$. We visualise these counts for dataset A3 as heatmaps in Figure 2.11, using the spatial distribution of spectra discussed in Section 2.3.2 and colouring pixels to indicate the count represented in $\mathfrak{d}^T\mathbb{X}$ for each spectrum. In Figure 2.11, we use light/bright colours to indicate pixels for which many of the DIPPS-features are present in the corresponding spectra, and dark/dull colours to indicate pixels for which many of the DIPPS-features are absent. Grey indicates spectra in which none of the DIPPS-features are present.

The strength of heatmaps such as those shown in Figure 2.11 is their interpretability — they provide interpretations for spatial regions to be that are of direct biological relevance. DIPPS-heatmaps allow for gradual differences between spatial regions to be represented — in contrast to the hard boundaries in a cluster membership. Furthermore, the values in the DIPPS-heatmaps can be directly interpreted in terms of the DIPPS-features, which correspond to peptide masses. For example, "between 5 and 10 DIPPS-features are present in a particular region". This strength

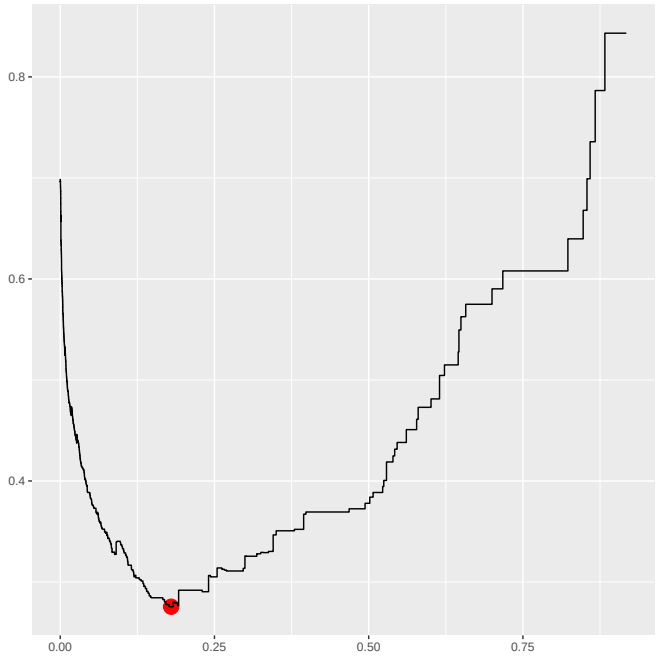


Figure 2.10: Plot of a on the x-axis versus $D_{cos}\left(\boldsymbol{c}, \boldsymbol{t}_{a^+}\right)$ on the y-axis, showing the cutoff a_*^+ in red where $\boldsymbol{c}$ is the centroid and $\boldsymbol{t}_{a^+}$ is defined in Definition 11 such that the subset of interest corresponds to the cancer cluster of the unsmoothed clustering result shown in Figure 2.6.

of DIPPS-heatmaps is highlighted by comparing to the relatively abstract interpretations of cluster memberships such as those of Figure 2.6. For example, "spectra in the cancer cluster are more similar to each other than to spectra from other clusters". We discuss the interpretations that can be made from such heatmaps later in this section, when we interpret the heatmaps of Figure 2.11. First, we discuss the application of the spatial smooth introduced in Section 2.5 to producing the smoothed results in Figure 2.6 and Figure 2.11.

The unsmoothed heatmap shown in Figure 2.11 appears "speckled" in places, and this speckling can be reduced by incorporating a spatial smooth. In Section 2.5 we suggest a spatial smooth that preserves the binary nature of the data. Dataset A3 was smoothed using Algorithm 2.2 with a smoothing parameter $\tau = 0.25$. An identical analysis, as discussed above, k-means clustering and DIPPS feature extraction, was performed on the smoothed data, resulting in the smoothed heatmap shown in Figure 2.11. Some things to notes about the results on the smoothed data include:

- The clustering is quite robust to smoothing, the two clusterings of Figure 2.6 differ in only 894 or $\sim 6\%$ of 14050 spectra. These small differences correspond to a similar reduction in 'speckling'.
- The 9 spectra that were empty in the raw data, as mentioned in Section 2.3.2, are no longer empty in the smoothed data.
- In the smoothed data $a_*^+ = 0.1926893$ for the cancer cluster, quite similar to $a_*^+ = 0.1831931$ in the initial analysis.
- There are 45 DIPPS-features in the smoothed data. All 45 show up in both the initial analysis, and the shifted-bin analysis mentioned earlier.

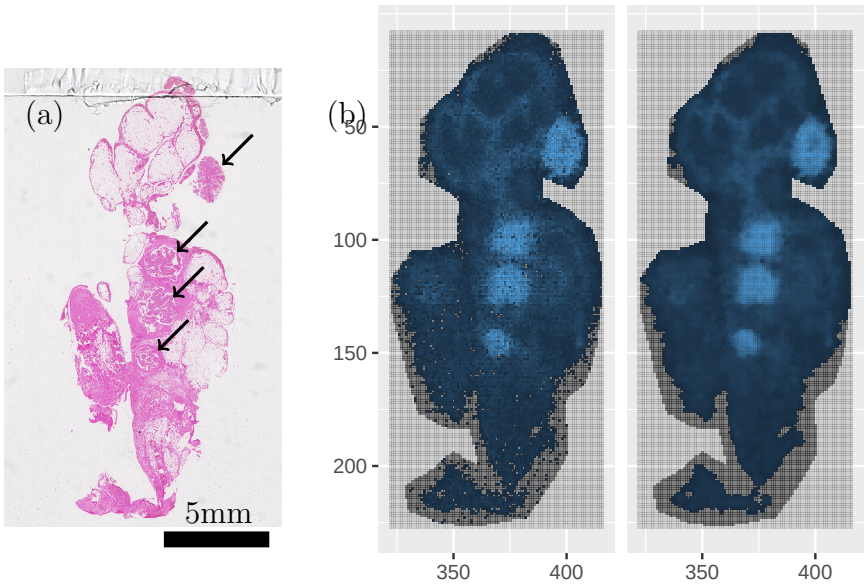


Figure 2.11: (a) H&E stained tissue section with arrows indicating the four visible primary tumours and (b) DIPPS heatmap for the cancer cluster in the unsmoothed data ($\tau = 0$, left) and the smoothed data ($\tau = 0.25$, right). In the unsmoothed data, the DIPPS-heatmap represents the sum of the 54 m/z bins with DIPPS $\geq a_*^+ = 0.1832$, and in the smoothed data the sum of the 45 smoothed ($\tau = 0.25$) m/z bins with DIPPS $\geq a_*^+ = 0.1927$.

The smoothed DIPPS-heatmap exhibits noticeably "sharper" edges and less "speckling" than the unsmoothed heatmap, while still displaying a very similar spatial distribution.

Comparing the heatmaps to the histology, all shown in Figure 2.11, the four main bright areas in the heatmap correspond well with the ovarian tumours, much like the cancer cluster of Figure 2.6. There are two less bright, but still distinguishable regions also of interest shown in the heatmaps of Figure 2.11:

- One "connects" between the top two cancer tumours, and extends up and to the left from them.
- The other is a separate node of brightness directly left of the second from bottom primary tumour.

When the tissue was considered by a pathologist, in addition to identifying the four main cancer tumours, two other areas of interest were noted, but could not be confirmed as cancerous tumours without further analysis. One of these additional areas corresponds to the second region noted above, the other was at the very bottom, and is not highlighted in the heatmaps of Figure 2.11. The region mentioned in the first point above was not highlighted by the pathologist as being potentially cancerous. This back and forth between MALDI-MSI data analysis and pathology is essential to making use of such data. In this case, the connecting region mentioned in the first point above has been identified as primarily connective tissue. One hypothesis that agrees with the pathology, and also explains why this region would be highlighted in the heatmaps of Figure 2.11 is that the tumours originated as connective tissue and may have retained some of the connective tissues characteristic molecular features. This hypothesis is further supported by the intermediate brightness of the regions between the three central tumours — that also corresponds to small areas of connective tissue. The area to the left has some similarity to the tumours and this is of

interest as it was identified as potentially cancerous by the pathologist, yet the area at the bottom which was also identified as potentially cancerous by the pathologist does not share this similarity to the tumours. It is possible that MALDI-MSI data could be used in combination with pathologists annotations in order to improve the sensitivity of histopathological annotations in the future. Furthermore the region at the bottom does *not* exhibit this similarity, and as such is clearly differentiable from the left area on the basis of the MALDI-MSI data — this demonstrates that multiple regions, equally 'in question' from the perspective of a pathologist, could potentially be distinguished by these molecular features. This is promising as it indicates that perhaps the use of MALDI-MSI data in combination with pathologists annotations based on staining and light microscopy could potentially not only improve sensitivity but also specificity of such histopathological annotations.

Chapter 3

Applications of DIPPS-based Feature Extraction

In Chapter 2 we introduced DIPPS-based feature extraction as a method for finding a set of DIPPS-features that are good positive indicators for a subset of interest in the data. In Section 2.6 we considered the application of this feature extraction method to the the ovarian cancer dataset A3 — in that application the subset of interest corresponded to cancer tissue in dataset A3, as obtained by k-means clustering. The DIPPS-based feature extraction approach is a general method however, requiring only binary data and a subset of interest. There is no limitation to MALDI-MSI data, or to using clustering to find such a subset. In this chapter we consider two applications of the DIPPS-based feature extraction approach introduced in Chapter 2. Both the applications we consider are MALDI-MSI applications as this is the focus of our work, but the DIPPS-based feature extraction approach can apply to any binary presence/ absence data, not only MALDI-MSI data. The first application we consider is a natural extension of the ovarian cancer work we began discussing in Chapter 2 and so uses clustering to find the subset of interest, but the second application does not require any analysis to find a subset of interest as the subset is already part of the experimental design. Brief descriptions of these two applications follow.

First, in Section 3.1 and with additional details in Appendix B, we extend the application considered in Section 2.6 to the ovarian cancer data of Section 1.5.1 in two ways: by applying the feature extraction to other tissue types, and by applying the feature extraction to the remainder of the ovarian cancer datasets discussed in Section 1.5.1, including both multiple datasets from the same patient and datasets from different patients. Ultimately we compare the sets of DIPPS-features extracted and thereby investigate within and between patient variability in order to demonstrate that within patient variability is less than between patient differences — meaning it is feasible to detect real between patient differences in these data. We have published this work (Winderbaum et al., 2015).

Second, in Section 3.2.2, we consider a different application of the DIPPS-based feature extraction approach, to the glycan data of Section 1.5.2. In these glycan data the goal is to demonstrate that glycans can be detected using MALDI-MSI through the use of an enzyme, PNGase F. In order to do this, two regions of tissue where used, one treated with PNGase F and one not. The untreated region is expected to exhibit no glycan signals and acts as a control group, so any glycans should be able to be detected by comparing these two groups. This natural separation of the data into two groups provides the subset of interest for the DIPPS-based feature extraction approach, and so no analysis is necessary to find the subset of interest

in this case, in contrast to the ovarian cancer application where we use k-means clustering to find the subsets of interest.

3.1 Comparing Ovarian Cancer Datasets

This section is organised in the following way. Initially I briefly describe the Jaccard distance and how it can be used to compare two sets of DIPPS-features. I suggest that this will provide a useful method for making the many comparisons we are interested in making between tissue types, datasets, and patients. Although the primary interest is the comparison of datasets from different patients, in order to do this in a meaningful way first the within-patient variability must be addressed. I will consider within-patient variability in Sections 3.1.2 - 3.1.4 by comparing results between datasets collected from the same patient. Having developed an understanding of the within-patient variability in Sections 3.1.2 - 3.1.4, I then consider comparisons between different patients in Section 3.1.5. Finally, I discuss my conclusions and the implications of these comparisons in Section 3.1.6. In Section 3.1.2 I consider results in some detail, as an example, but in Sections 3.1.3 - 3.1.5 I focus on the most important and interesting results. A more detailed discussion of these results is included in Appendix B.

3.1.1 Jaccard Distance for Comparing Datasets

Gorzolka and Walch (2014) have shown that comparisons between tissue samples is complicated by tissue (and tumour) heterogeneity, to the point where even samples from the same patient can appear to be very different. In order to detect meaningful differences between patients despite high within-patient variability, it is crucial to take tissue heterogeneity into account. The ability to take tissue heterogeneity into account is the primary advantage of MALDI-MSI, as it has the potential to separate data from different tissue types within a single tissue sample. In Section 2.3.3 we demonstrated that tissue types can be separated by clustering in the ovarian cancer data. In Section 2.6.2 we demonstrated how a DIPPS-based feature extraction approach could be used to find a set of positive indicators which we call DIPPS-features, for a particular subset of the data such as a cluster or tissue type. Here I briefly describe the Jaccard distance and how it can be used to compare two sets of DIPPS-features.

Definition 13. Jaccard Distance: *is a measure of the dissimilarity between two sets S_1 and S_2,*

$$D_{Jac}(S_1, S_2) = 1 - \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}. \tag{3.1}$$

I use $|S_1|$ to denote the size of the set S_1, or its cardinality. Note that other set-theoretic measures could be used here, $|S_1 \cup S_2| - |S_1 \cap S_2|$ for instance, but these would all be largely equivalent for our purposes and the Jaccard distance is a commonly used measure of dissimilarity between sets across a wide variety of disciplines — see Cross and Sudkamp (2002); Jaccard and Jacoby (2010); Leydesdorff (2008). When S_1 and S_2 are sets of DIPPS-features, corresponding to m/z bins or variables in the MALDI-MSI data, the Jaccard distance is directly interpretable as the proportion of DIPPS-features unique to one of the two sets being compared. The simple form of the Jaccard distance is useful. If a particular similarity or dissimilarity is of interest, identifying the variables that contribute to that similarity/

dissimilarity is straightforward, and so targeted follow-up experiments can be easily designed. In the remainder of Section 3.1 we apply the DIPPS-based feature extraction approach of Section 2.6.2 using the heuristic cutoff of Definition 12 for each of the tissue types/ clusters in each of the ovarian cancer datasets introduced in Section 1.5.1. We will then make extensive use of the Jaccard distance to compare the sets of DIPPS-features resulting from this feature extraction approach.

3.1.2 Detailed Comparisons Within Patient A

In Chapter 2 dataset A3 was used extensively in illustrative and exploratory analyses. A3 corresponds to a single section or 'slice' of an excised tissue block from patient A. The datasets A1, A2 and A4 correspond to different sections or 'slices' of the same tissue block as A3. As all four of these sections come from the same tissue block, it is expected that they should be very similar. As such any reliable analysis method should produce similar results on these sections. For example, clustering to separate tissue types should show visually similar patterns. How similar/ different results are between these datasets can give an impression of the within-patient variability as well as the reproducibility of both the technology, and the method of analysis. Furthermore A1 and A2 correspond to consecutive sections of tissue, and thus would be expected to be even more similar to each other. In this section I will introduce these other patient A datasets, and apply the ideas introduced in Chapter 2 to them. I present clustering results in Figure 3.1, and apply the feature extraction approach of Section 2.6.2 using the heuristic cutoff of Definition 12 to obtain a set of DIPPS-features for each cluster. I then visualise the similarities/ differences between these sets of DIPPS-features using the Jaccard distance, Definition 13, in Figure 3.2. This visualisation allows for the sets of DIPPS-features associated to each of the 16 clusters shown in Figure 3.1 to be compared, both within dataset and between datasets.

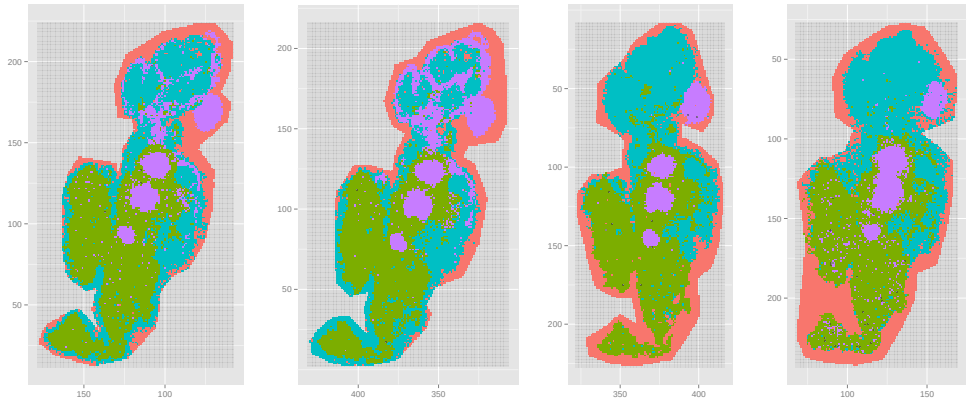


Figure 3.1: Spatial maps showing the cluster membership produced by 4-means clustering with the cosine distance on the binary binned data for 4 datasets, from left to right: A1, A2, A3 and A4. Clusters are identified with colours, and roughly correspond to the tissue-types cancer, adipose, stroma, and off-tissue.

Figure 3.1 shows the cluster-membership produced by 4-means clustering using the cosine distance on the binned binary data for the four datasets: A1, A2, A3 and A4. As mentioned above datasets A1 and A2 come from consecutive sections of tissue, meaning that discounting any distortion caused during sectioning, these

two sections of tissue should be approximately $6 - 10\,\mu$m apart in the direction orthogonal to the sectioning — roughly one tenth of a pixel, so we would expect them to be very similar. Similarly, A3 and A4 come from nearby sections, with only two or three sections between them. We can see from Figure 3.1 that in all four datasets the cluster analysis results in clusters that correspond spatially to the four broadly different tissue types present in the tissue — cancer, adipose, stroma, and off-tissue. There is a noticeable difference in the cancer clusters of datasets A1 and A2 as compared to A3 and A4 in that the connective areas at the top of the images shown in Figure 3.1 are included in the cancer clusters of datasets A1 and A2, but not of A3 and A4. These connective areas are the same as mentioned in Section 2.6.3. In Section 2.6.3 we hypothesised that this region was similar to the cancer in A3 due to the tumour having grown out of this connective tissue, and these clustering results support this hypothesis — the similarity is potentially even more evident in datasets A1 and A2 in which this region is clustered together with the primary tumours, indicating spectra from this region are sufficiently similar to spectra from the primary tumours to be clustered together. Also, for datasets A1 and A2 some of the off-tissue spectra in the lower Y-coordinate values that are closer to the tissue are included in the adipose cluster. In datasets A3 and A4 the adipose clusters do not extend to these spectra, and there are more off-tissue spectra total, particularly in dataset A4. To summarise, there is broad similarity in the clustering results, but when considered in more detail, results from consecutive sections are more similar than results from non-consecutive sections — this is to be expected.

Figure 3.2 shows the 16×16 symmetric distance matrix resulting from pairwise Jaccard distance comparisons between the 16 sets of DIPPS-features, each corresponding to one of the 16 clusters shown in Figure 3.1 — four from each of the four sections. This matrix of comparisons is visualised in Figure 3.2 using colour to show values of the Jaccard distance — dark colours indicating small values close to zero, i.e. similarity, and light colours indicating large values close to one, i.e. dissimilarity. Each pixel in the 16×16 grid shown in Figure 3.2 corresponds to a single pairwise comparison. For example, the far top right pixel of Figure 3.2 corresponds to the comparison of the off-tissue and cancer clusters both from dataset A4. In this example, a set of DIPPS-features is found for each of these two clusters using the feature extraction method described in Section 2.6.2 and the heuristic cutoff of Definition 12. These two sets of DIPPS-features are then compared using the Jaccard distance, and the Jaccard distance between the two sets determines the colour of the pixel. In this example, the pixel is very light coloured meaning the Jaccard distance between them is close to one, that they are dissimilar, and more specifically that the two sets being compared do not have very much overlap between them.

Remarks on Figure 3.2:

- The dark diagonals are very clear, reflecting the broad agreement between datasets. For simplicity let us interpret clusters as tissue types, despite small inconsistencies between the two. Then we can interpret these dark diagonals as showing that the variables which characterise any particular tissue type in one dataset, often also characterise that tissue type in other datasets. Because the Jaccard distance is a set comparison measure, dark pixels, small values, or similar sets indicate a large intersection between the two sets being compared. In this case the sets correspond to DIPPS-features found to characterise particular tissue types in particular datasets, and these diagonals represent comparing the characterising DIPPS-features for the same tissue type across different datasets. The fact these diagonals are dark show that variables identified as DIPPS-features for a particular tissue type in one dataset often are

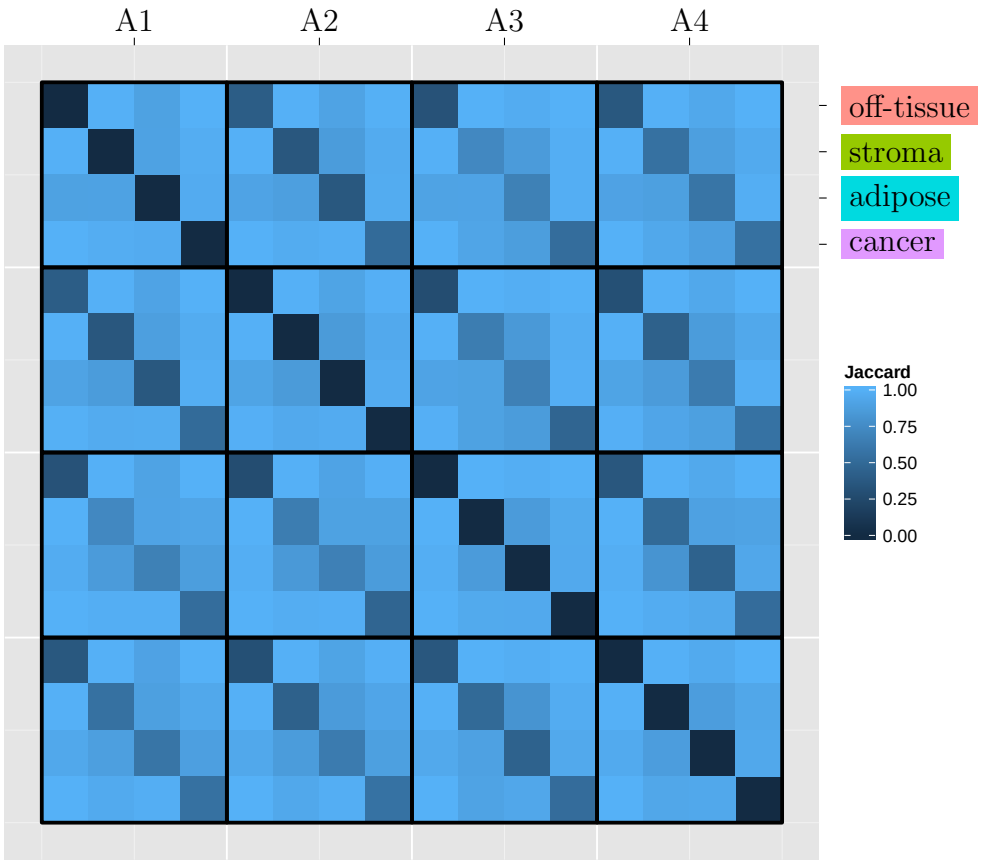


Figure 3.2: Image representing the Jaccard distance comparisons of Section 3.1.1 of the cluster memberships of Figure 3.1. A set of DIPPS-features is found for each of the 16 clusters shown in Figure 3.1 using the feature extraction approach discussed in Section 2.6.2 and the heuristic cutoff of Definition 12. The image shown above represents pairwise Jaccard distances between these sets of DIPPS-features. Black lines separate datasets, with the four pixels within each black divisor corresponding to the four clusters for that dataset. So the main block diagonal represents within-dataset comparisons, and the rest represent between-dataset comparisons.

identified as DIPPS-features for the same tissue type in other datasets as well — i.e. these sets have a large intersection. This essentially establishes the reproducibility of the DIPPS-feature extraction method for reliably identifying the same (or similar) variables important to the characterisation of tissue types across different datasets from the same patient.

- The similarity between the cancer cluster of A1 and A2 to the adipose cluster of A3 and A4 noted during the discussion of Figure 3.1 above is reflected in Figure 3.2, but using the results of Figure 3.2 we could, if we were interested, now identify the variables responsible for this similarity. This allows for more indepth interpretation than was possible from the clustering results of Figure 3.1 alone.

- The only off-diagonal entry that is consistently and noticeably darkened between all the datasets is the comparison of adipose and stroma clusters, which could indicate that of all the tissue types, these are the least well separated, and we will see further evidence for this when we consider the clustering results on the other patients in Sections 3.1.3 and 3.1.4.

3.1.3 Summary of Comparisons Within Patient B

In this section we consider four datasets from patient B, another ovarian cancer patient, as described in Section 1.5.1. I will introduce the four datasets much like I did for patient A in Section 3.1.2, by considering the clustering results. However for patient A I discussed in detail the Jaccard distance comparisons as well as the clustering results, and although these detailed comparisons are included in Appendix B they are omitted here for brevity.

Figure 3.3 shows cluster-membership much like Figure 3.1, but for the four datasets from patient B. In Section 3.1.2 I equated clusters to tissue types, and mentioned that this was a slight simplification as there are small discrepancies between the two that can be seen by comparing the cluster memberships to the H&E stained tissues sections. In the interest of brevity I do not include the H&E images here, instead describing any notable discrepancies between the clustering results and the tissue types, as these will be relevant when interpreting the comparisons to follow. Broadly we are interested in tissue types, and I often refer to the clusters by their associated tissue types, i.e. cancer, stroma, adipose, and off-tissue. However, when discussing discrepancies between tissue types and clusters, in order to clarify the distinction I use the colour to refer specifically to a cluster and not its associated tissue type, i.e. purple, green, cyan, or salmon. The purple clusters of datasets A1 and A2 in Figure 3.1 containing some connective tissue regions as well as the cancerous primary tumours as discussed in Section 3.1.2 is an example of such a discrepancy between clusters and tissue types. It is important to discuss these discrepancies as they explain many of the features in the comparisons that would otherwise appear to be artefacts. Only once these discrepancy-caused effects are understood can the remaining comparisons be interpreted and overall conclusions be made — I discuss such overall conclusions in Section 3.1.6.

As far as discrepancies between tissue types and clusters as shown in Figure 3.3 go, the main point to note is that in dataset B1 the cancer and stroma clusters are well separated, but in the other three datasets the clustering broadly grouped these two tissue types together in the same cluster. This leaves an 'extra' cluster in these three datasets, because these two tissue types are both included in a single cluster. In B2 and B4 the off-tissue region is split between the salmon and green clusters.

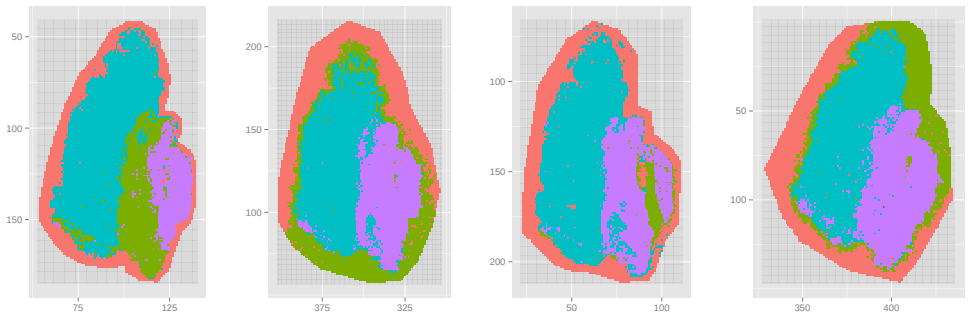


Figure 3.3: Spatial maps showing the cluster membership produced by 4-means clustering with the cosine distance on the binary binned data for 4 datasets, from left to right: B1, B2, B3 and B4. Clusters are identified with colours, and roughly correspond to the tissue-types cancer, adipose, stroma, and off-tissue. The exception to this rough correspondence between clusters and tissue types is that the green clusters of datasets B2 and B4 correspond to off-tissue regions.

In B3, the green cluster forms a small area between cancer areas. It is also useful to note that the adipose clusters across all four datasets agree well, both with each other and with the histology.

In order to reduce discrepancies between clustering results and tissue types as determined by histology, the clustering results could be improved by using more sophisticated clustering methods and fine-tuning parameter choices and data cleaning steps that preceded the clustering. However, even without these refinements these tissue types can be separated by individualising the choice of number of clusters to use for each patient, as it would appear the stroma is more difficult to separate from the cancer in patient B. In Winderbaum et al. (2015), we use 3-means clustering for patient B and this achieves much better agreement between cluster membership and tissue types. However the objective here is to demonstrate the usefulness of the DIPPS feature extraction in exploratory analyses and the fact that these clustering results *do not* perfectly reflect the tissue types actually helps to highlight the usefulness of DIPPS — it is not necessary for the clustering to separate the tissue types perfectly in order for us to obtain useful interpretations using this approach. The fact that we use the exact same clustering algorithm on all the datasets also helps to simplify the presentation of these results, allowing us to focus our discussion on the interpretation of results without needing to carefully justify many parameter choices.

The clustering results for B1 correspond best to tissue morphology, and so I will compare the other datasets clustering results to those for B1.

3.1.4 Summary of Comparisons Within Patient C

Similarly to Section 3.1.3, here I introduce clustering results for four datasets from patient C — shown in Figure 3.4. The more detailed Jaccard distance comparisons of these clusters associated DIPPS-features are included in Appendix B.

Remarks on the clustering results of Figure 3.4:

- C2 and C4 show a noticeable number of empty spectra, indicated by grey pixels. This could indicate a problem in overall data quality. Reasons that could account for this above-normal number of empty spectra include: inef-

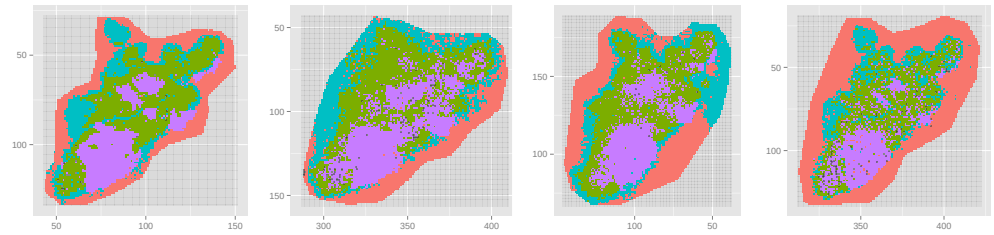


Figure 3.4: Spatial maps showing the cluster membership produced by 4-means clustering with the cosine distance on the binary binned data for 4 datasets, from left to right: C1, C2, C3 and C4. Clusters are identified with colours, and roughly correspond to the tissue-types cancer, adipose, stroma, and off-tissue.

fective antigen retrieval, digestion, or matrix deposition during sample preparation, instrumentation issues during data acquisition, and even artefacts in the peak-picking step. For our purposes, the key thing to remember is that these datasets could potentially have lower data quality compared to the other datasets.

- In C2 and C3 the cyan clusters extends out into the off-tissue region. This is similar to the cyan clusters in datasets A1 and A2 as shown in Figure 3.1. One explanation for these clusters including off-tissue regions is that there could be peptides that are mobilised during sample preparation which move off the tissue. The fact that clusters from different datasets, and even different patients, show this spread into off-tissue regions from adipose tissue suggests the possibility that some adipose specific peptides may be particularly susceptible to being mobilised. This is a concern that could be further investigated and such investigation could ultimately result in improvements to sample preparation methods for acquiring accurate and spatially resolved MALDI-MSI data in the future.

Aside from the two points above, cluster agreement both between datasets and to tissue types is good.

3.1.5 Between Patient Comparisons

A comprehensive comparison of all four clusters from each of the 12 datasets introduced is of interest, and is included in Appendix B. However we are particularly interested in the cancer comparisons, as similarities therein could potentially yield common markers that could be used for early detection. Dissimilarities in the cancer comparisons could potentially lead to markers for diagnosis, prognosis, or prediction of response to treatment and be used for the individualisation of treatment plans. Due to our particular interest in cancer we consider the cancer comparisons specifically in Figure 3.5. Figure 3.5 shows the pairwise Jaccard distances between the 12 sets of DIPPS-features, each corresponding to one of the 12 cancer clusters shown across Figures 3.1, 3.3, and 3.4 — one set of cancer cluster characterising DIPPS-features from each of the four sections from each of the three patients. Here we discuss these results in a broad sense, and we provide some more details in Section 3.1.6 and Appendix B.

The first thing to note in Figure 3.5 is that within-patient similarities are stronger than between-patient comparisons. This is encouraging as tissue heterogeneity,

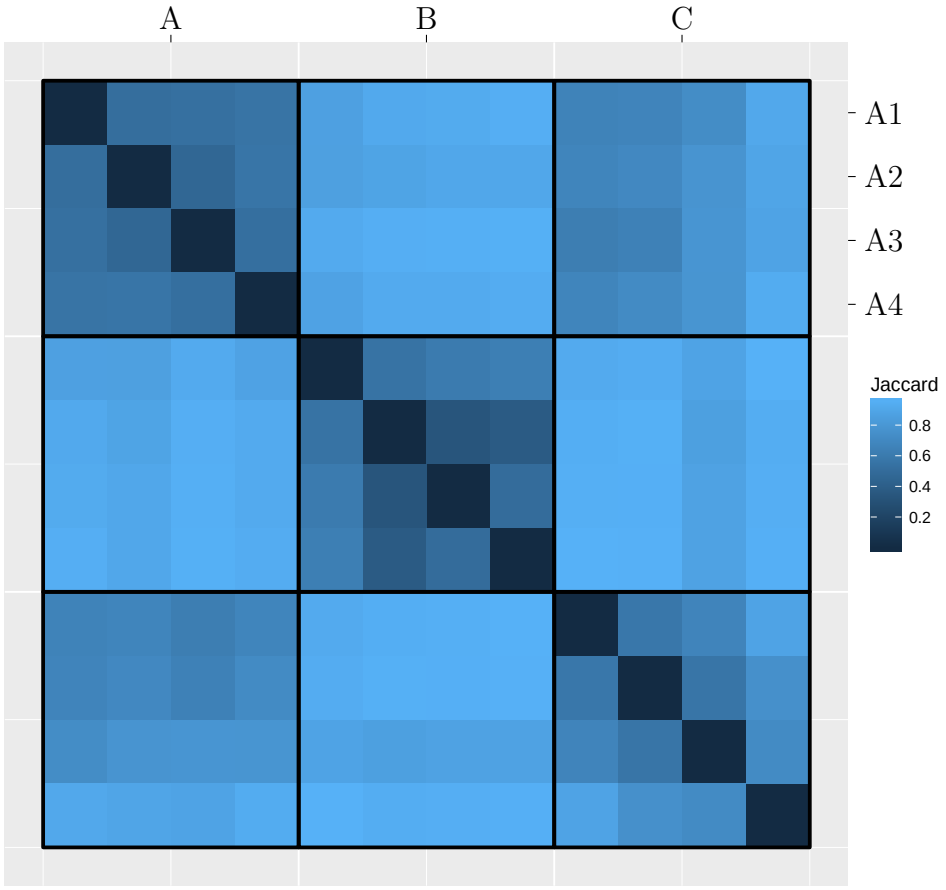


Figure 3.5: Image representing the pairwise Jaccard distances of Section 3.1.1 between the 12 sets of DIPPS-features corresponding to each of the 12 cancer clusters shown in Figures 3.1, 3.3, and 3.4 — one set of cancer cluster characterising DIPPS-features from each of the four sections from each of the three patients. Black lines seperate patients, similarly to Figure B.3 — infact the results shown here are a subset of those in Figure B.3, corresponding to every fourth pixel, i.e. those corresponding to cancer clusters.

which we are hopefully at least partly accounting for by using our clustering results, can provide challenges in the analysis of MALDI-MSI data (Gorzolka and Walch, 2014). After we note that we can separate within-patient variability from between-patient variability by visually observing that the main block diagonal in Figure 3.5 is darker than the off-diagonal, the next observation from Figure 3.5 is also clear — there appears to be strong similarity between patients A and C that is absent in comparisons with patient B. There are also a few more minor points:

- Dataset C4 has lower similarities to all other datasets overall, and this could be explained by the potentially lower quality of this dataset implied by the 'speckling' noted earlier.

- Dataset B1 has lower within-patient similarity than any other dataset — this is likely to be because dataset B1 is alone amongst the patient B datasets in separating the cancer tissue well, the other three datasets include significant amounts of stromal tissue in their purple clusters.

The second point above leads back to the main observation that the cancer clusters of patient B are notably different from those of patients A and C. Initially one might think this could be due to the inclusion of non-cancer stroma tissue in three of these purple clusters (B2, B3, and B4), however note that the cancer cluster of dataset B1 (in which the cancer cluster is much better separated) does not show much higher cross-patient similarity. This indicates that a more likely explanation for patients B's cancer cluster demonstrating dissimilarity to the cancer clusters of patients A and C is actually that there is a difference between these patients cancers at a molecular level. We discussed this result that the cancer of patient B seems to differ from the cancer of patients A and C with our biochemist collaborators at the APC, and after acquiring more detailed descriptions of the histopathology of the tissues from these patients came to understand that the cancer tissue of patient B is largely necrotic, and it could be that we are detecting blood-specific masses that characterise this necrosis. In short, the back and forth process between data analysis and biology/histopathology produced a real, biologically significant and consistent, difference between these patients' cancers.

3.1.6 Conclusions

In Chapter 2 we demonstrated that powerful interpretations can be made from extending clustering results with a DIPPS-based approach to identifying sets of positive indicators for each cluster. In Section 3.1.1 we then discussed how the Jaccard distance could be used to compare these sets of positive indicators. In Sections 3.1.2-3.1.5 and Appendix B we then demonstrated how meaningful interpretations can be obtained by considering these comparisons. These interpretations range from quality control checks to biologically relevant clues that warrant further investigation via LC-MS and immunohistochemistry for identification and validation at the protein level. We also demonstrated that meaningful conclusions could be drawn despite the variations and artefacts of the clusterings — infact, the variation in clustering results enriched the conclusions drawn from the Jaccard distance comparisons, despite making them more complicated. Some particular examples of interpretations resulting from the Jaccard distance comparisons follow:

- There are 7 m/z bins that characterise the off-tissue regions in all 12 datasets — these could be further investigated as potential contaminants, matrix peaks, or artefacts of the peak-picking algorithm. If they could be confirmed as matrix

peaks, for example, they could then be excluded in future studies. Until further validation is carried out on these peaks, they should certainly be considered sceptically if they came up as relevant in any other analyses.

- Similarly, there are 16 *m/z* bins that characterise the green clusters in both B2 and B4 (which correspond to off-tissue regions), but characterise none of the off-tissue clusters in any of the datasets. These *m/z* values could be further investigated as potentially delocalised peptides (if they are confirmed to be peptides), or some form of localised contamination (fingerprints, saliva, etc.) specific to these sections. 7 of these *m/z* bins also characterise the cyan clusters in C2 and C3, which include both tissue and off-tissue regions, possibly supporting the hypothesis that these could be delocalised peptides. Given the sample preparation steps, there should be only very minimal homogenisation and so if these *m/z* bins could be validated as mobilised peptides, this would indicate directions in which the sample preparation steps could be improved. Either way, these masses warrant further investigation.

- Of the 98 variables that characterise the purple clusters in B2, B3, and B4 (which contain both cancer and stroma tissue), 67 also characterise the cancer cluster of B1 (where cancer and stroma are well separated). However, of these 67 *m/z* bins, 60 also characterise the stroma cluster of B1. This highlights the difficulty in separating the cancer from the stroma in this patient, even in dataset B1 where the clustering succeeded in doing so. This is most likely due to how the cancer grew out of the stroma tissue and so is still molecularly quite similar to it. The 7 variables that characterise the cancer but not the stroma in patient B could be further investigated as potential biomarkers.

- The intersection of all 12 sets of cancer-characterising DIPPS-features contains exactly one *m/z* bin. That *m/z* bin is centred at 1628.75. This *m/z* value matches to a peptide that was identified as originating from the protein 'Heterogeneous nuclear ribonucleoprotein A1' (ROA1) in follow-up LC-MS. The identity of this peptide was validated using *in situ* MS/MS. This ROA1 protein has been shown to be of interest in the past (Lee et al., 2010; Chen et al., 2010). The identity of this protein has since been validated by immunohistochemistry, and is one of the proteins of interest noted in Winderbaum et al. (2015). The *m/z* bin centred at 1628.75 is one of the 7 bins mentioned in the dot point above as characterising all four cancer clusters of patient B, but not the stroma cluster of dataset B1.

- The other 6 *m/z* bins of the 7 mentioned above are also interesting, as they do not characterise any other cancer clusters (in either patient A or C) and as such could be further investigated as potential markers that distinguish the cancer of patient B from that of patients A and C — possibly even for necrotic tissue in general.

It is possible to investigate many of the results discussed above in more depth. For example, analyses with shifted bin locations could be considered. Also, more detail could be included in the investigation of individual *m/z* bins — for example, there are (in addition to those mentioned in the first point above) another 15 *m/z* bins which are positive indicators for off-tissue in 11 (of 12) datasets.

However, the emphasis of this thesis is to demonstrate that bioinformatic approaches as suggested here can produce substantial and meaningful results with relative ease. For the bioinformatician implementing these methods, it is important

to realise that there are many small improvements that can be made to squeeze every little bit of information out of their data in any particular case, but as most of the conclusions from these types of exploratory analyses will involve follow up validation studies, this is not necessarily the goal, and is not what we focus on here.

3.2 Exploratory Analysis of the Murine N-glycan Data

As discussed in Section 1.6, in order to represent peaklist MALDI-MSI data in the standard statistical paradigm of 'variables' and 'observations' we need to discretise the m/z domain and thereby group peaks by m/z into variables. Up to here we have used the data-independent binning approach to this, as described in detail in Appendix A. However in some cases, such as the N-glycan data that we will consider here, a data-dependent approach is more appropriate. Data-dependent methods overcome some of the disadvantages of the data-independent binning approach, for example binning can split peaks that ought to be grouped into the same variable due to its arbitrary bin locations, and data-dependent approaches are much less likely to do this. The disadvantage of data-dependent methods is, as mentioned in Section 2.1, that extending analyses to multiple datasets is much less natural. As the interest in the N-glycan data is not to compare multiple datasets, but rather to simply explore the one dataset, a data-independent approach to discretisation of the m/z domain is appropriate.

This section is organised as following. First, we introduce the data-dependent discretisation method we will use in Section 3.2.1. Then in Section 3.2.2 we consider the application of the DIPPS-based feature extraction discussed in Section 2.6.2 to the N-glycan data of Section 1.5.2.

3.2.1 Tolerance Clustering

Discretisation can be thought of as clustering of one dimensional data — m/z. The result of discretisation is essentially a 'cluster membership', with each peak belonging to a 'cluster' or variable. In the case of binning, these 'clusters' correspond to bins.

Here we introduce one of the simplest approaches to data-dependent discretisation — tolerance clustering. Tolerance clustering is a fairly simple concept, and has been used to discretise peaklist MALDI-MSI data in the past, for example see Gustafsson et al. (2012). Tolerance clustering can be thought of as the process of forming equivalence classes defined by the equivalence relation by which two values are equivalent if and only if the absolute difference between them is $\leq \tau$ for some given grouping tolerance τ. To be precise, given a set of real numbers m_k, which in the context of peaklist MALDI-MSI data would be the m/z values of peaks, and a grouping tolerance $\tau > 0$ the relevant equivalence relation is

$$C_\tau = \Big\{(m_k, m_{k'}) \quad s.t. \quad |m_k - m_{k'}| \leq \tau \Big\}. \tag{3.2}$$

The equivalence classes defined by Equation 3.2 are then the clusters or peakgroups or variables, as in Definition 14.

Definition 14. Tolerance clustering: *Given a set of real numbers m_k a grouping tolerance $\tau > 0$ and the equivalence relation C_τ as in Equation 3.2 tolerance clustering results in the set of group labels c_k assigning each m_k to a cluster such*

that $c_k = c_{k'}$ *if and only if there exists a sequence* $k_1, k_2, \ldots, k_n$ *for which* $k_1 = k$, $k_n = k'$ *and* $(m_{k_i}, m_{k_{i+1}}) \in C_\tau$ *for all* $i = 1, 2, \ldots, (n-1)$.

Reasonable values for the grouping value τ will depend on the application, and choosing an appropriate value will require some experimentation in any given case, but in peptide and glycan MALDI-MSI peaklist data, values of τ around 0.1 are often reasonable as the difference in m/z between features is often at least 1, and the mass error is typically smaller than 0.1. Other applications in which the mass error is greater, such as protein MS, may require larger values of τ.

Both tolerance clustering and binning have the advantage that the number of clusters does not need to be specified *a priori*, as it does with a clustering method such as k-means. However there are also differences between tolerance clustering and binning, as mentioned previously. For example, tolerance clustering produces clusters such that if two values are within τ of each other they are guaranteed to be in the same cluster. Binning cannot make such a guarantee — two values separated by τ can be put into different bins if the boundary between two bins happens to fall between them. The interpretability awarded by this guarantee is one of the main advantages of tolerance clustering. Following are several other points worth mentioning about tolerance clustering in the context of MALDI-MSI data. Note that in this context, we interpret clusters as 'peakgroups' as the values we are clustering are peaks m/z values.

- Tolerance clustering has a tendency to produce a number of peakgroups that are very small, containing as few as a single peak. It is quite similar to binning in the sense that binning also produces many almost empty variables. Typically further analyses are robust to the removal of these low-occurrence peakgroups (variables), as we demonstrated for binning in Section 2.6.1. Removing these low-occurrence peakgroups can improve both computational speed and ease of interpretation by reducing the number of variables involved.

- The choice of tolerance τ is important — results can be very sensitive to choice of τ if τ is chosen to be too large. If τ is too large, tolerance clustering will group values that should be kept separate. If the dataset is sufficiently large, or if enough datasets are combined for a single analysis, this grouping of values that should be kept separate can occur regardless of τ. If τ is too small, tolerance clustering can split values that should be kept together, compounding the problem discussed in the first dot point above. Peakgroups are typically very well resolved in MALDI-MSI particularly for high-quality internally calibrated datasets such as those we consider. In our experience, a tolerance of $\tau = 0.1$ tends to avoid both these issues for sufficiently small datasets — less than two million peaks or so. In larger datasets, say 10 or 15 million peaks, tolerance clustering will often begin to fail regardless of choice of τ unless the rate of erroneous and high mass-error peaks can be limited somehow. In these datasets, we recommend using either a data-independent approach with fixed cluster width such as binning, or a more sophisticated data-dependent discretisation that can take into account peak density, such as the `DBSCAN` algorithm proposed by Ester et al. (1996) or more specifically its deterministic variant `DBSCAN*` (Campello et al., 2013, Section 3).

- Tolerance clustering can exhibit artefacts — particularly when there is a single outlying peak halfway between two groups of peaks — causing the tolerance clustering to group the two together, when they should be kept separate.

Many of the pitfalls of tolerance clustering mentioned in the dot points above can be avoided by more advanced data-dependent clustering methods for discretisation. However, in many cases it is possible to, by carefully choosing a tolerance τ, avoid all these issues in any given MALDI-MSI dataset. As mentioned above, a tolerance of $\tau = 0.1$ will often work, with almost no problems. If possible, using tolerance clustering in this way is advantageous due to its simple interpretability.

In conclusion, although tolerance clustering improves on some of the limitations of binning, it also shares many of the limitations of binning. Some of these limitations of the tolerance clustering method can be further improved upon by using more advanced clustering algorithms, as mentioned above, however in many cases these limitations can be minimised by careful selection of tolerance τ. When these limitations can be minimised by careful selection of τ then tolerance clustering has the advantage that the resulting peakgroups are quite easily interpretable, as mentioned above.

The glycan data of Section 1.5.2 is such a case for which careful choice of tolerance τ can minimise the limitations of the tolerance clustering method, and as such we apply tolerance clustering to these data in Section 3.2.2. Improvements could be made by using more sophisticated clustering approaches, however the purpose of the N-glycan was to serve as a proof-of-principle and not to squeeze as much information from the data as possible, and so tolerance clustering is more than sufficient for exploring the N-glycan data.

3.2.2 Using the DIPPS in the Context of Glycan Data

The glycan data of Gustafsson et al. (2015), as introduced in Section 1.5.2, consists of 202667 peaks from 11014 spectra collected from two regions of interest, one treated with PNGase F in order to release N-linked glycans, the other a control that should contain no glycan signals. The objective in this glycan experiment was to demonstrate that this sample preparation approach can successfully detect glycans using MALDI-MSI and preserve the spatial distributions of these glycans in the process. We use this application to illustrate the usefulness of the DIPPS feature extraction approach in a context different to the ovarian cancer data originally used to illustrate it in Section 2.6.2. In these glycan data, there is a region of tissue that has been treated with PNGase F, which cleaves glycans and makes them detectable by MALDI-MS. A separate control region, not treated with PNGase F, is also present, and so potential glycan signals should be identifiable by occurring in the PNGase F treated regions but not in the control region. This natural partitioning of the data into two groups, control group and PNGase F treated group, provides the subset of interest necessary for calculating the DIPPS. Note that the DIPPS-based feature extraction approach applies to any subset of interest, that can originate from many different sources. In contrast to the ovarian cancer data analysis of Section 3.1, where k-means clustering was used to find the subsets of interest, here the subset of interest is inherent to the design of the experiment and no analysis is necessary in order to find it. The aim of the analysis to follow is to find a shortlist of candidate glycan masses by comparing the two regions using the DIPPS, and show that at least some of them have spatial distributions matching the histology of the tissue.

Samples were analysed by LC-MS in parallel, also with and without PNGase F treatment — and glycans were identified in the PNGase F treated samples. The shortlist of glycan candidates we produce from the analysis of the MALDI-MSI data will then be compared to the glycans identified by LC-MS in order to further support the identities of the shortlist masses as glycans. We will also compare the spatial

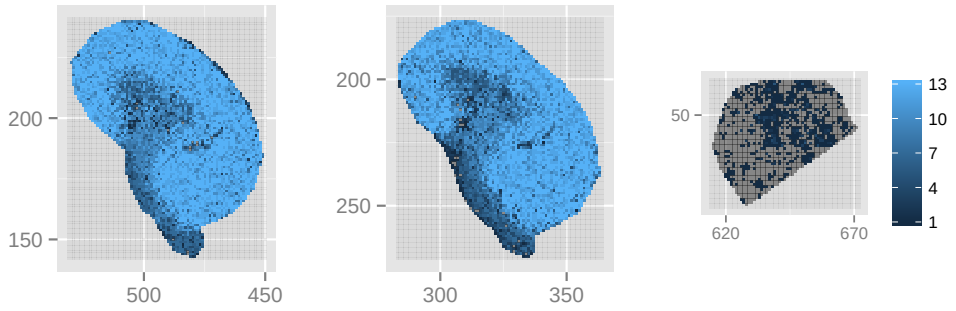


Figure 3.6: DIPPS heatmaps showing the sum (count) of the occurrence in the 13 peakgroups with DIPPS $\geq a_*^+ = 0.586$ in the PNGase F treated region (left and centre) and the control region (right). Analogous to the heatmaps of Figure 2.11.

distributions of these candidate masses with tissue histology, as it is of interest if tissue-type specific glycans can be found.

Tolerance clustering using a grouping tolerance of $\tau = 0.1$ on the glycan data produces 850 peakgroups, i.e. 850 unique values of the group labels c_k in Definition 14. As mentioned above, many of these peakgroups contain very few peaks, similar to the distribution shown in Figure 2.7 and discussed in Section 2.6.1 — 90% of these peakgroups contain fewer than 100 peaks. As mentioned above, one concern when using tolerance clustering is that an outlying peak lying halfway between two peakgroups can cause the otherwise separate peakgroups to be grouped together when they ought to be kept separate. A good sanity check for such unwanted groupings is to consider the range of each peakgroup — the difference between the minimum and maximum m/z in the peakgroup. In the glycan data, the maximum such range after tolerance clustering with a tolerance of $\tau = 0.1$ is 0.737 Da. An m/z difference of 1 Da is the smallest difference between features we expect to be able to resolve in these data, and so the fact that the maximum range is less than 1 Da reassures us that any resolvable features differing by 1 Da have not been combined into a single peakgroup. If we had peakgroups with ranges significantly above 1 Da, it might be worth considering a lower tolerance τ or either a data-independent discretisation or a more sophisticated clustering method that can take into account peak density. In the case of the glycan data, tolerance clustering with $\tau = 1$ is sufficient to resolve any features differing by at least 1 Da and so we continue with this choice.

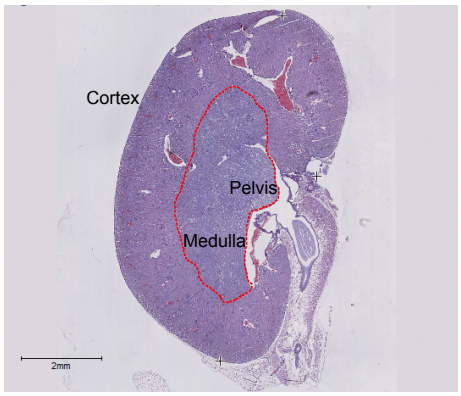


Figure 3.7: H&E stain of a section of murine kidney used in the glycan experiment. Red annotation shows the border between the outer cortex region, and the inner medulla/pelvis region.

We are interested in finding variables, that is peakgroups produced by the tol-

erance clustering discussed above, containing peaks in the PNGase F treated region but not in the control region. Such variables would be positive indicators for the PNGase F region, and so we choose the subset of interest to be the PNGase F treated region when we calculate the DIPPS for each variable. This choice is analogous to us choosing the subset of interest to be the cancer tissue in the analysis of the ovarian cancer data in Section 2.6.2. Using the heuristic cutoff of Definition 12, the DIPPS ranking includes 13 peakgroups with DIPPS $\geq a_*^+ = 0.586$. The spatial distribution of the sum (count) of occurrence in these 13 peakgroups is shown in Figure 3.6 as a DIPPS heatmap, and Figure 3.7 shows the histology of a typical mouse kidney section for comparison, with the major components annotated including the outer cortex region, and the inner medulla/pelvis region. The log-intensity of the 13 peakgroups of Figure 3.6 are shown individually in Figure 3.8 and Figure 3.9(a). Of these 13 peakgroups, the spatial distributions appear to be either uniformly distributed across the kidney (Figure 3.8(a), (i), (j), and (l)), cortex-specific (Figure 3.8(e-h,k), Figure 3.9(a)), or somewhere inbetween (Figure 3.8(b-d)). This is not representative of glycan distributions in this tissue, as the cortex makes up a disproportionate amount of the tissue, and DIPPS will rank glycans higher if they cover a higher proportion of the PNGase F treated tissue. Although the heuristic cutoff was useful in order to be able to make the comparisons between ovarian cancer datasets in Section 3.1, in the context of the glycan data we are interested in finding as many potential glycan masses as possible, and so it makes sense to extend the list of masses we consider to include more than the top 13. Spatial distributions for the log-intensity of the peakgroups ranked $1-12$, $13-24$, and $25-36$ as ranked by highest DIPPS are shown in Figures 3.8, 3.9, and 3.10 respectively. The associated m/z location and DIPPS for each of these top 36 ranked peakgroups are included in the figure captions. The PNGase F treated region includes two sections of kidney, and the figures show only one for simplicity as replication is good between the two sections — this can be seen in the similarity between the spatial patterns of the centre and left images of Figure 3.6. Note that we stop at the 36th ranked peakgroup and don't go further because by visually inspecting Figure 3.10 in comparison to Figure 3.9 and Figure 3.8 we can see that most of the peakgroups ranked $25-36$ shown in Figure 3.10 are low-occurrence and many of them are not spatially localised in comparison to more highly ranked variables. If we were to continue and consider variables ranked below 36, this trend only continues, and so the more lowly ranked variables are not of interest. Considering the extended list of peakgroups shown in Figure 3.9 and Figure 3.10, we can note several more interesting spatial distributions:

- There are two peakgroups, Figure 3.9(h) and (j), that exhibit a distinct cortex-specific spatial distribution concentrated in the centre of the kidney close to the medulla but absent in the outer cortex. This distinct cortex-specific spatial distribution seems to indicate there are at least two regions of the cortex which differ in their glycan composition — very interesting!

- There are also three distinct spatial distributions specific to the medulla/pelvis region:

  - Figure 3.9(f), Figure 3.10(b), and (g) show spatial distributions focused in a particular subset of the medulla, in a shape spread along the short axis of the kidney.
  - Figure 3.10(e) and (l) show spatial distributions also focused in a different subset of the medulla, in a shape spread along the long axis of the kidney.

– Finally, Figure 3.10(c) and (i) show highly specific spatial distributions that seem to correspond to a blood vessel visible in the H&E stain of Figure 3.7.

To conclude, distinct spatial distributions can certainly be observed in these data, most of which correspond well with the histology.

As can be seen from Figure 3.8, there are many peakgroups with very high DIPPSs, which is encouraging for the experiment, as this most likely indicates that glycans have been successfully extracted from the tissue and measured. In order to further support this conclusion, we can match the m/z values to the LC-MS results, and verify their identity by *in situ* MS/MS. Of the top 26 peakgroups as ranked by DIPPS, 16 were successfully matched with LC-MS and assigned potential identifications. These 16 matches are shown in Table 3.2 with estimated mass errors and proposed glycan compositions and structure. A legend for the symbols used to draw the proposed glycan structures in Table 3.2 is provided in Table 3.1.

Gustafsson et al. (2015) also demonstrate that different spatial distributions can be discovered in these data without consideration of particular masses — specifically they successfully separate the two cortex regions, and one region covering the whole medulla/pelvis region. Regardless, the conclusion is that these glycan data are a proof-of-principle that it is possible to measure consistent spatial distributions of glycans in FFPE tissue using MALDI-MSI.

The analysis of the glycan dataset also provides some insight into the heuristic of Definition 12. It is interesting to note that the heuristic cutoff is effectively 0.5 in this case as the 14th peakgroup, i.e. Figure 3.9(b), has a DIPPS of 0.41. Peakgroups with high DIPPS are expected to be glycans and so are not expected to occur in the control region at all. The DIPPS is the difference of two proportions of occurrence, one from the subset of interest which in this case is the PNGase F region, the other from the complement which in this case is the control region. As glycans are expected to have a proportion of occurrence of zero in the control region, the only factor influencing the value of their DIPPS is their proportion of occurrence in the PNGase F treated region. Following from the discussion at the end of Section 2.6.2, the heuristic of Definition 12 attempts to minimise the distance between the centroid of the subset of interest and the DIPPS-template representing the most highly ranked variables by DIPPS. The most highly ranked peakgroups in the context of the glycan data should all be glycans, and for these peakgroups the DIPPS reduces to the proportion of occurrence in the PNGase F treated region. So if we begin with the empty set, the corresponding DIPPS-template will be the vector of zeros. As we add more peakgroups, working down from the most highly ranked by DIPPS we switch zeros into ones in the corresponding elements of the DIPPS-template, until we reach a peakgroup with a DIPPS below 0.5 — the 14th peakgroup in the glycan data. Switching the 14th zero to a one will increase the distance between the centroid and the DIPPS-template rather than decreasing it, and so the heuristic DIPPS-threshold selects the top 13 peakgroups. The above interpretation of the heuristic is a slight oversimplification, as it implicitly assumes that the centroid of the subset of interest is the mean. When using the cosine distance, this interpretation is only valid if each spectrum has the same number of peaks and therefore their binary representations have the same length. In practice observations do not all have the same number of peaks, but this interpretation is still useful to explore how the heuristic of Definition 12 behaves.

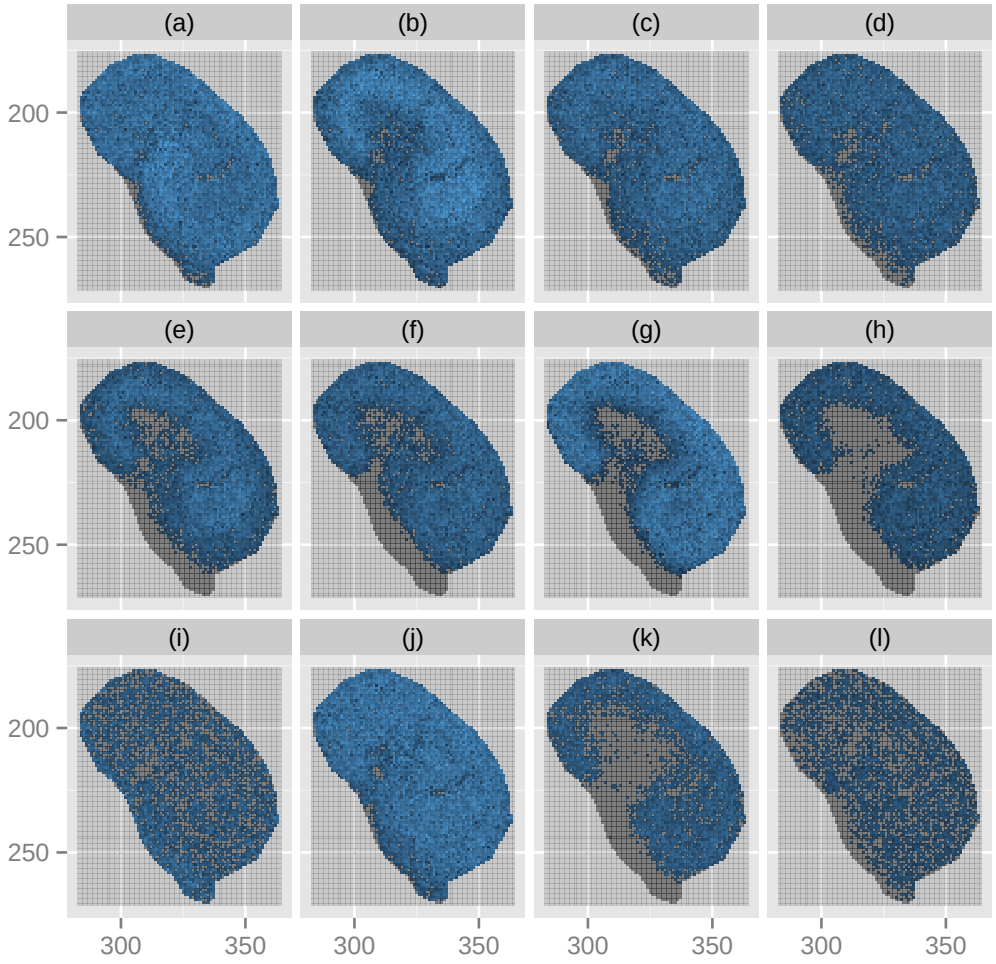


Figure 3.8: Spatial plots with log-intensity of peaks shown using colour, bright colours corresponding to high log-intensities dark to low log-intensities and grey indicating the absence of peaks. Each subplot shows a peakgroup, and the peakgroups shown are those ranked $1 - 12$ in decreasing order of DIPPS. For each peakgroup an abundance weighted mean m/z was calculated by averaging the m/z of all peaks in the peakgroup, weighting based on their SNR. Shown are the peakgroups with SNR-weighted mean m/z of:

(a) 1257.47 and DIPPS 0.97.

(b) 1905.7 and DIPPS 0.93.

(c) 1743.64 and DIPPS 0.92.

(d) 1581.59 and DIPPS 0.87.

(e) 1850.73 and DIPPS 0.83.

(f) 1996.79 and DIPPS 0.83.

(g) 2304.91 and DIPPS 0.82.

(h) 2158.84 and DIPPS 0.66.

(i) 822.94 and DIPPS 0.66.

(j) 1419.53 and DIPPS 0.64.

(k) 917.35 and DIPPS 0.62.

(l) 1809.69 and DIPPS 0.6.

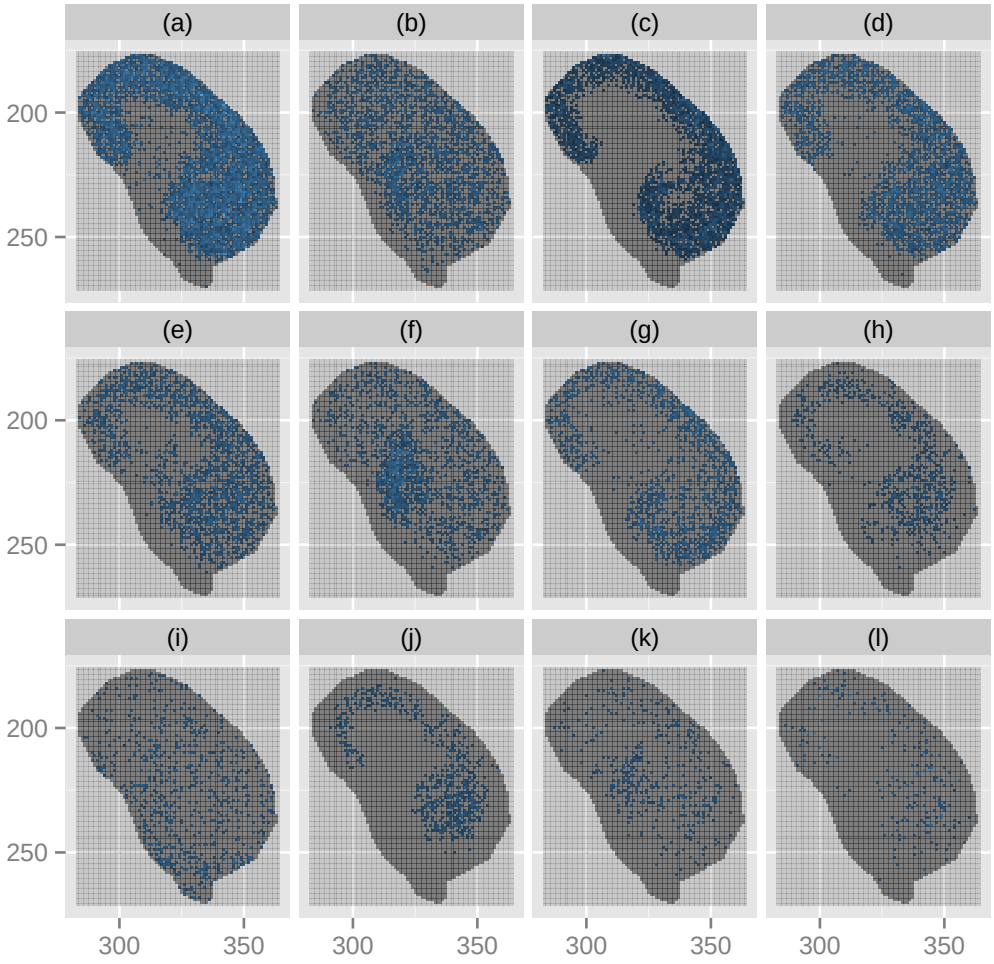


Figure 3.9: Spatial plots with log-intensity of peaks shown using colour, bright colours corresponding to high log-intensities dark to low log-intensities and grey indicating the absence of peaks. Each subplot shows a peakgroup, and the peakgroups shown are those ranked $13 - 24$ in decreasing order of DIPPS. For each peakgroup an abundance weighted mean m/z was calculated by averaging the m/z of all peaks in the peakgroup, weighting based on their SNR. Shown are the peakgroups with SNR-weighted mean m/z of: .

(a) 1079.41 and DIPPS 0.59.

(b) 1663.63 and DIPPS 0.41.

(c) 2816.12 and DIPPS 0.41.

(d) 933.34 and DIPPS 0.39.

(e) 1485.59 and DIPPS 0.34.

(f) 1688.66 and DIPPS 0.28.

(g) 1095.4 and DIPPS 0.26.

(h) 2012.77 and DIPPS 0.15.

(i) 1042.58 and DIPPS 0.13.

(j) 2067.75 and DIPPS 0.12.

(k) 1647.64 and DIPPS 0.09.

(l) 1282.51 and DIPPS 0.05.

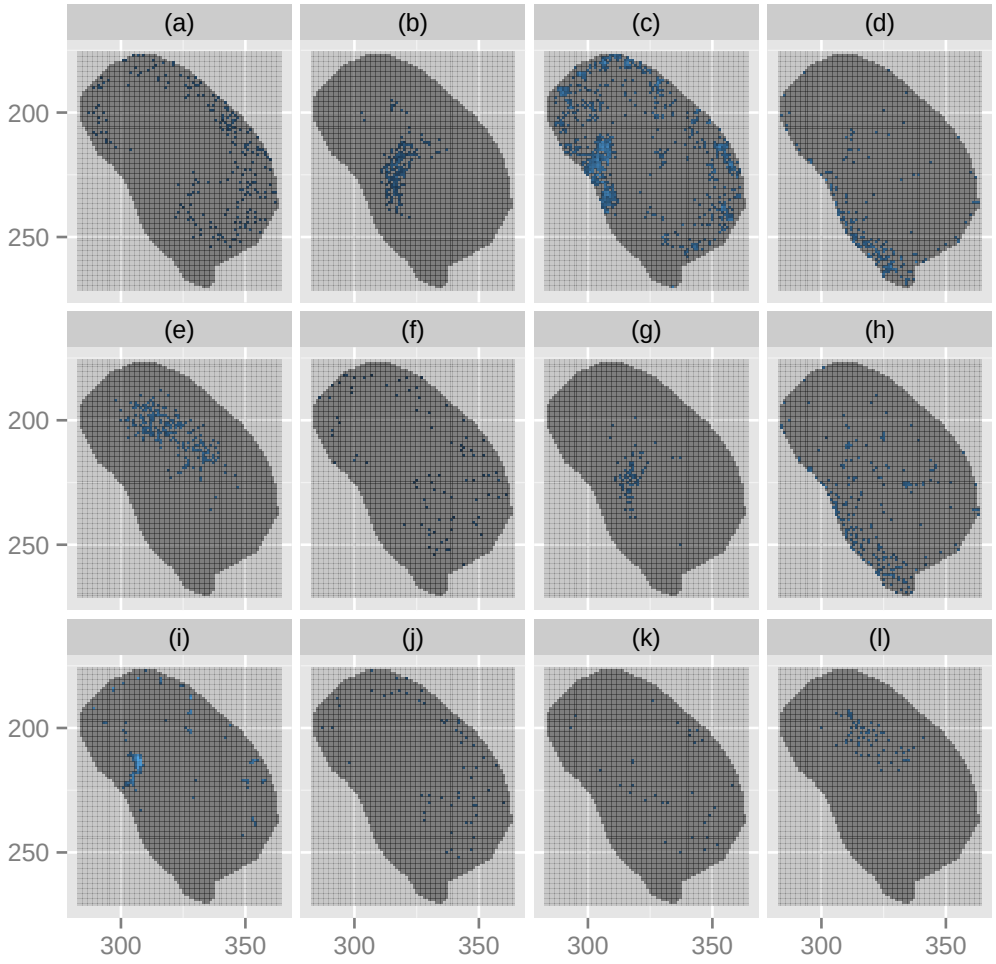


Figure 3.10: Spatial plots with log-intensity of peaks shown using colour, bright colours corresponding to high log-intensities dark to low log-intensities and grey indicating the absence of peaks. Each subplot shows a peakgroup, and the peakgroups shown are those ranked $25 - 36$ in decreasing order of DIPPS. For each peakgroup an abundance weighted mean m/z was calculated by averaging the m/z of all peaks in the peakgroup, weighting based on their SNR. Shown are the peakgroups with SNR-weighted mean m/z of: .

(a) 2507.99 and DIPPS 0.05.

(b) 2321.78 and DIPPS 0.04.

(c) 1330.8 and DIPPS 0.03.

(d) 806.98 and DIPPS 0.03.

(e) 931.56 and DIPPS 0.02.

(f) 2815.1 and DIPPS 0.02.

(g) 1992.7 and DIPPS 0.02.

(h) 829.02 and DIPPS 0.02.

(i) 1231.41 and DIPPS 0.01.

(j) 2142.84 and DIPPS 0.01.

(k) 1971.73 and DIPPS 0.01.

(l) 909.56 and DIPPS 0.01.

Symbol	Monomer
	N-acetylglucosamine
	Fucose
	Mannose
	Galactose
	Glucose

Table 3.1: Glycan Notation Legend

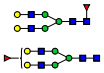
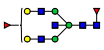
Figure	LC-MS/MS $[M + Na^+]^+$ calculated	mass error (ppm)	Proposed composition	Proposed structure	Xu et al. (2012)
3.8(a)	1257.41	50.2	$(Hex)_2 + (Man)_3(GlcNAc)_2$		1257.42
3.8(j)	1419.47	43.9	$(Hex)_3 + (Man)_3(GlcNAc)_2$		1419.48
3.9(e)	1485.53	40.9	$(HexNAc)_2(Deoxyhexose)_1 + (Man)_3(GlcNAc)_2$		1485.53
3.8(d)	1581.53	35.8	$(Hex)_4 + (Man)_3(GlcNAc)_2$		1581.53
3.9(k)	1647.57	39.7	$(Hex)_1(HexNAc)_2(Deoxyhexose)_1 + (Man)_3(GlcNAc)_2$		1647.59
3.9(b)	1663.57	37.1	$(Hex)_2(HexNAc)_2 + (Man)_3(GlcNAc)_2$		1663.58
3.9(f)	1688.61	31.6	$(HexNAc)_3(Deoxyhexose)_1 + (Man)_3(GlcNAc)_2$		1688.61
3.8(c)	1743.57	41.7	$(Hex)_5 + (Man)_3(GlcNAc)_2$		1743.58
3.8(l)	1809.63	33.2	$(Hex)_2(HexNAc)_2(Deoxyhexose)_1 + (Man)_3(GlcNAc)_2$		1809.64
3.8(e)	1850.65	43	$(Hex)_1(HexNAc)_3(Deoxyhexose)_1 + (Man)_3(GlcNAc)_2$		1850.67
3.8(b)	1905.63	35.4	$(Hex)_6 + (Man)_3(GlcNAc)_2$		1905.63
3.9(h)	2012.71	31.5	$(Hex)_2(HexNAc)_3(Deoxyhexose)_1 + (Man)_3(GlcNAc)_2$		2012.72
3.9(j)	2067.67	37.2	$(Hex)_7 + (Man)_3(GlcNAc)_2$		2067.69
3.8(h)	2158.77	34.4	$(Hex)_2(HexNAc)_3(Deoxyhexose)_2 + (Man)_3(GlcNAc)_2$		2158.78
3.8(g)	2304.83	34.44	$(Hex)_2(HexNAc)_3(Deoxyhexose)_3 + (Man)_3(GlcNAc)_2$		2304.83
3.9(c)	2816.01	37.4	$(Hex)_3(HexNAc)_4(Deoxyhexose)_4 + (Man)_3(GlcNAc)_2$		

Table 3.2: Matched masses between the *in situ* MALDI acquisition and the LC-MS/MS. For proposed structures and more details, see Gustafsson et al. (2015). A legend for the symbols used to draw the proposed structures is provided in Table 3.1.

Chapter 4

Methods for Classification

In this chapter we will introduce methods for the classification of MALDI-MSI data. In Chapter 5 we will apply these classification methods to the TMA data introduced in Section 1.5. Classification involves the construction of a 'classification rule', which is used to assign an observation to a class. In Section 4.1 we introduce three methods from the literature as well as introducing the concept of Cross Validation (CV) as an approach for judging the performance of a classification rule. One of the challenges to classifying MALDI-MSI data is their high-dimensional nature, and one approach to addressing this challenge is to use some form of variable reduction prior to classification. In Section 4.3 we introduce two approaches to variable reduction.

In Section 4.1 and Section 4.3 we introduce some existing methods from the literature. In contrast, in Section 4.2 and Section 4.4 we introduce original contributions. In Section 4.2 we discuss some of the challenges to classifying MALDI-MSI data, including the conflicting advantages of data-dependent vs. data-independent discretisation of the m/z domain, as discussed in Section 3.2. Ultimately, we discuss our approach to preprocessing MALDI-MSI data prior to classification. In Section 4.4 we introduce a new way to reduce unwanted variability in MALDI-MSI data, which although initially computationally difficult is made tractable by the derivation of an analytic form for the inverse of a particular class of matrices. The derivation of the analytic form for this matrix inverse is included in Appendix C. These sections discuss very different aspects of the same problem — classification of MALDI-MSI data. In Section 4.5 we summarise how these different aspects connect and allow for such classification to be performed and improved, before moving on to Chapter 5 where we apply these ideas to the TMA data of Section 1.5.3.

4.1 Classification and Cross Validation

Classification, sometimes called discriminant analysis, can be subdivided into two steps:

- Constructing a classification rule capable of assigning a class label to an observation. The construction of the rule is done on the basis of data with known class membership, sometimes called 'training' data. This construction step is often referred to as the 'training' or 'learning' step.
- Applying a rule to assign a class label to an observation (or observations). This step can be further subdivided into one of two cases:
  - Applying the rule to observations of known class membership in order to assess the performance of the rule — sometimes called 'testing'.

– Applying the rule to observations of unknown class membership for which a real-world decision needs to be made — i.e. prediction.

In Section 4.1.1 we consider the 'testing' case, in which the performance of a classification rule is assessed by applying it to data with known class membership. In Sections 4.1.2, 4.1.3, and 4.1.4 we consider the first step above — constructing a classification rule on the basis of data with known class membership. In Sections 4.1.2 and 4.1.3 we introduce two classical approaches that originate with Fisher (1936) and remain canonical in the current classification literature. Finally, in Section 4.1.4 we introduce a more modern approach to classification specifically developed by Marron et al. (2007) to address challenges encountered in the classification of high-dimensional data.

It should be noted that there are a plethora of approaches to classification, as discussed in more detail in Section 1.6.2, and here we consider only a very limited selection. We will restrict attention to linear classification approaches, but it should be noted that many more non-linear alternatives exist. In general, linear methods are easier to interpret, particularly in the context of high-dimensional data. Therefore the simpler linear methods are often favoured over non-linear alternatives in High-Dimension Low Sample Size (HDLSS) contexts. We will also be restricting attention to two-class classification, as in the endometrial data the interest is to discriminate between patients with positive and negative Lymph Node Metastasis (LNM) status. It should be noted that many of these classification methods have natural generalisations to classification problems involving more than two classes. Some of our notation will hint at these generalisations but we focus on the two-class case as this is the case relevant to the data we consider. As we restrict attention to the linear two-class case for classification, we introduce some notation here specific to this case. We use this general notation to compare between the classification approaches we consider in Sections 4.1.2, 4.1.3, and 4.1.4. Let $\mathbb{X}$ be a $d \times n$ data matrix of n observations with known class labels coded as -1 or $+1$. All the rules we will consider use the data $\mathbb{X}$ and the associated class labels to 'train' a rule by finding a $d \times 1$ vector $\boldsymbol{d}$ and a scalar β. This rule then assigns class label $\tau(\boldsymbol{x})$ to a $d \times 1$ observation $\boldsymbol{x}$, which can be either a column of $\mathbb{X}$ or a new observation, in the following way:

$$\tau(\boldsymbol{x}) = \begin{cases} +1 & \text{if} \quad \boldsymbol{d}^T\boldsymbol{x} + \beta > 0 \\ -1 & \text{if} \quad \boldsymbol{d}^T\boldsymbol{x} + \beta < 0 \end{cases} . \tag{4.1}$$

Note that Equation 4.1 does not address the case when $\boldsymbol{d}^T\boldsymbol{x} + \beta = 0$, and in this unlikely case we do not assign a class label to $\boldsymbol{x}$. The different classification approaches we discuss in Sections 4.1.2, 4.1.3 and 4.1.4 each essentially constitute different choices for $\boldsymbol{d}$ and β.

To illustrate the intuition behind the notation of Equation 4.1, we present a short example application here. In this example we will apply LDA, as introduced in Section 4.1.2, to a subset of Anderson's iris data. Fisher (1936) introduced the canonical LDA, and demonstrated its usefulness by applying it to Anderson's iris data. The role of LDA became so fundamental in the field of classification that the iris data presented in the original paper has come to be known famously as 'Fisher's iris data'. Although Fisher is justifiably credited with the development of the canonical LDA method, 'Fisher's iris data' on the other hand should perhaps more accurately be known as 'Anderson's iris data' due to the contribution of Anderson (1935) towards quantification of the morphological variation amongst the iris species of the Gaspé peninsula, as Fisher (1936) himself acknowledged. Anderson's iris data consist of 4 measurements on 50 iris flowers from each of three different

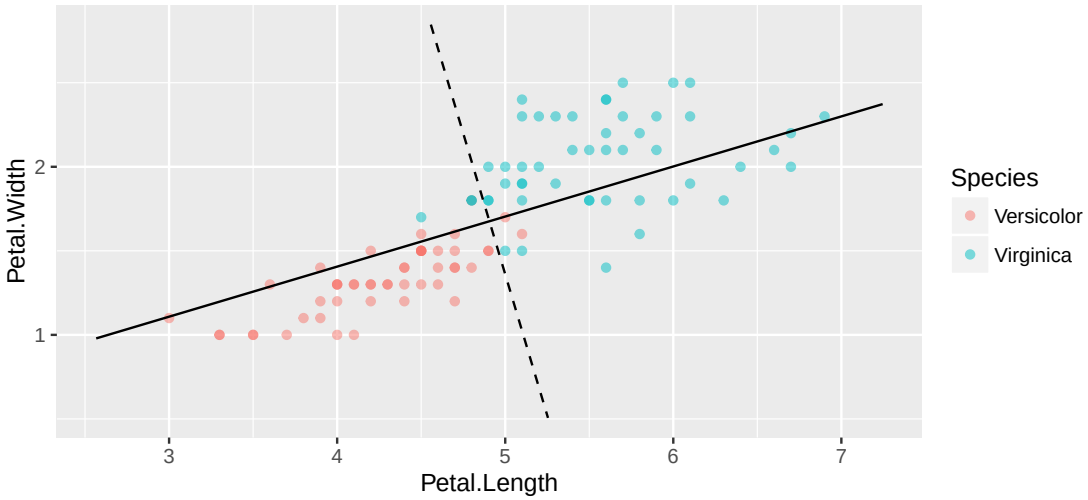


Figure 4.1: Petal width vs. petal length, both measured in cm, for a subset of Fisher's iris data, specifically the Versicolor and Virginica irises. The discriminating direction, $\boldsymbol{d}_{\text{LDA}}$, as found by LDA is shown as a solid black line. The separating hyperplane is shown as a dashed black line, and represents the cutoff value, β_{LDA}, such that observations on one side of the hyperplane (on one side of β_{LDA} on the '$\boldsymbol{d}_{\text{LDA}}$ axis') will be classified as one class and observations on the other side will be classified as the other class.

species. We will consider two of the measurements, petal length and petal width, from observations of two species, Versicolor and Virginica. Figure 4.1 shows these data, plotting the two measurements against each other and using colour to distinguish the two species. When applied to these data, the training step of LDA results in $\boldsymbol{d} = \boldsymbol{d}_{\text{LDA}}$ and $\beta = \beta_{\text{LDA}}$, which are visually represented in Figure 4.1 with solid and dashed black lines respectively. $\boldsymbol{d}_{\text{LDA}}$ can be thought of as a direction or line which 'best separates' the two classes, as determined by LDA. β_{LDA} can then be though of as the point (or perpendicular hyperplane — in this case a line) along the line/ direction defined by $\boldsymbol{d}_{\text{LDA}}$ that best separates the two classes — again, with 'best' being determined by LDA. The differences between linear classification methods can be thought of as different approaches to determining the meaning of the word 'best' in this context. As we can see from Figure 4.1, if we were to train and test the LDA classification rule on these data we would misclassify 5 irises — three Virginica would be misclassified as Versicolor(two dots are overlapped close to the dashed black line) and two Versicolor would be misclassified as Virginica.

4.1.1 Misclassification and Cross Validation

Once a classification rule has been constructed on the basis of some $d \times n$ data $\mathbb{X}$ with known class membership, it is of interest to assess its performance. One method for assessing its performance is to use the rule to assign a class label to each of the observations of $\mathbb{X}$, whose class membership are known, and count how many have been assigned the incorrect class label. I will use the term 'misclassification' to refer to this count, but in the literature it is sometimes referred to as 'classification error' or 'misclassification rate'.

When attempting to assess the performance of a classification rule, particularly in a situation where $n < d$, the misclassification rate can be misleading due to

over-fitting effects. An effective approach to addressing the issue of over-fitting is to use two separate datasets — a 'training' set, and a 'testing' set. In practice, data collection can often be prohibitively expensive and access to large sample sizes is often not possible, particularly for rare diseases. Due to these limitations, we are motivated to find a compromise somewhere inbetween using separate 'training' and 'testing' datasets and using misclassification rate. Such a compromise would 'make the most' of a small dataset better than splitting it into two separate datasets, and would be less prone to over-fitting effects as compared to simple misclassification.

For $N \leq n$, N-fold CV is an approach to finding such a compromise. N-fold CV can be thought of as a sequence of steps:

- Construct N non-empty n-index subsets $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N$ such that they partition $\{1, 2, \ldots, n\}$. These n-index subsets represent a partition of the observations in the data, See Definition 5 on n-index subset notation. Usually, $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N$ are constructed to be as close to equal size as possible.

- Construct N classification rules τ_i for $i = 1, 2, \ldots, N$ where τ_i is constructed or 'trained' on the basis of the subset

$$\bigcup_{j \neq i} \mathcal{C}_j$$

  of the data and the associated known class labels.

- Assign a class label to each observation, using τ_i to assign a class label to observations in $\mathcal{C}_i$ for each i. As the $\mathcal{C}_i$ partition the data, this process will assign each observation a unique class label. These assigned class labels can then be compared with the known (assumed to be true) class labels, and the number of observations whose true class labels disagree with their assigned class labels is called the N-fold misclassification.

The largest possible N is the number of observations n, and this special $N = n$ case is called Leave-One-Out (LOO) CV because in this case the above steps amount to each observation being left-out and a rule trained on the basis of the remaining data, and tested on the left-out observation. Although the most computationally intensive, LOO CV is appropriate for small sample sizes as it maximises the number of observations used in the construction of each rule, while still 'testing' each rule on an observation not used in its construction. LOO misclassification (n-fold misclassification) will be the main statistic by which we compare the performance of the various classification approaches we consider.

4.1.2 Fisher's Linear Discrimination Analysis

Although there are other linear classification methods, I will use the relatively generic term LDA to refer specifically to Fisher's LDA as described in this section. LDA is described in detail by Koch (2013, Section 4.3). The motivation behind LDA is intuitive — in order to separate our classes, Fisher (1936) suggests we aim to maximise the between-class variability, and minimise the within-class variability.

Let $\mathbb{X}$ be a $d \times n$ data matrix whose columns correspond to observations of known class membership. Let $\bar{X}$ denote the mean of the columns of $\mathbb{X}$, and let $\mathbb{X}^{[\nu]}$ to denote the submatrix of $\mathbb{X}$ consisting of the columns in class ν. Let $\bar{X}^{[\nu]}$ to denote the mean of the columns of $\mathbb{X}^{[\nu]}$ and let n_ν denote the number of observations in class ν (i.e. the number of columns of $\mathbb{X}^{[\nu]}$). Let $\bar{\bar{X}}$ denote the mean of the $\bar{X}^{[\nu]}$s and let $S^{[\nu]}$ denote the sample covariance matrix of $\mathbb{X}^{[\nu]}$.

We define the two matrices $\hat{B}$ and $\hat{W}$ as

$$\hat{B} = \sum_{\nu} \left(\bar{X}^{[\nu]} - \bar{\bar{X}}\right)\left(\bar{X}^{[\nu]} - \bar{\bar{X}}\right)^T \tag{4.2}$$

and

$$\hat{W} = \sum_{\nu} S^{[\nu]} \tag{4.3}$$

respectively. $\hat{B}$ of Equation 4.2 is the sample covariance matrix of the class means, representing between-class variability. $\hat{W}$ of Equation 4.3 is the sum of within-class covariance matrices, and represents the within-class variability.

LDA solves the optimisation problem of finding the direction (unit length) vector that maximises the between-class variance while minimising the within-class variance of the projected data (projected into that direction). It turns out that the optimal direction vector $\boldsymbol{v}$ is the eigenvector $\boldsymbol{d}_{\text{LDA}}$ associated to the largest eigenvalue of the matrix:

$$\hat{W}^{-1}\hat{B}. \tag{4.4}$$

Koch (2013, Section 4.3) includes a proof that the eigenvector $\boldsymbol{d}_{\text{LDA}}$ is the solution to this optimisation problem.

It is important to note that calculating $\boldsymbol{d}_{\text{LDA}}$ requires that $\hat{W}$ be invertible and in HDLSS cases, i.e. $n < d$, this is not possible. I present a more precise discussion of the conditions when $\hat{W}$ cannot be invertible in Section 4.1.3.

Also, note that $\boldsymbol{d}_{\text{LDA}}$ is only unique up to sign, so if we use the symbols '+' and '−' to denote the two classes $+1$ and -1 respectively, we choose the sign of $\boldsymbol{d}_{\text{LDA}}$ such that

$$\boldsymbol{d}_{\text{LDA}}^T \bar{X}^{[+]} > \boldsymbol{d}_{\text{LDA}}^T \bar{X}^{[-]}.$$

The LDA classification rule τ_{LDA}, constructed from $\mathbb{X}$ and associated class labels, is of the general form of Equation 4.1 and assigns the class label to a $d \times 1$ observation $\boldsymbol{x}$

$$\tau_{\text{LDA}}(\boldsymbol{x}) = \begin{cases} +1 & \text{if} \quad \boldsymbol{d}_{\text{LDA}}^T \boldsymbol{x} - \boldsymbol{d}_{\text{LDA}}^T \bar{\bar{X}} > 0 \\ -1 & \text{if} \quad \boldsymbol{d}_{\text{LDA}}^T \boldsymbol{x} - \boldsymbol{d}_{\text{LDA}}^T \bar{\bar{X}} < 0 \end{cases}. \tag{4.5}$$

4.1.3 Naive Bayes

'Naive Bayes (NB)' refers to an approach to modifying an existing classification method and is not a classification method itself. I use the term 'NB' to refer specifically to the NB variant of LDA which I describe in this section. NB modifies LDA in a way that allows it to function when $\hat{W}$ of Equation 4.3 is not invertible. As mentioned in Section 4.1.2, requiring that $\hat{W}$ of Equation 4.3 be invertible is problematic when $n < d$. Specifically, if we let κ denote the number of classes,

$$\operatorname{rank} \hat{W} \leq \sum_{\nu} \min\left(n_\nu - 1, d\right) \quad \text{simplifies to} \quad \operatorname{rank} \hat{W} \leq n - \kappa \quad \text{when} \quad \max_{\nu} n_\nu - 1 \leq d. \tag{4.6}$$

The inequality of Equation 4.6 means that if $n - \kappa < d$ then $\hat{W}$ is guaranteed to be singular.

The NB variant of LDA essentially makes the 'naive' assumption that variables are independent, and from this assumption it follows that their covariance should be zero. In practice, instead of calculating the direction vector from the matrix $\hat{W}^{-1}\hat{B}$ as in LDA, the direction vector for the NB variant $\boldsymbol{d}_{\text{NB}}$ is instead calculated as the eigenvector corresponding to the largest eigenvalue of the matrix

$$(\operatorname{diag} \hat{W})^{-1}\hat{B}. \tag{4.7}$$

In the case that the variables are infact independent, LDA reduces to NB.

Similarly to LDA we choose the sign of $\boldsymbol{d}_{\text{NB}}$ such that

$$\boldsymbol{d}_{\text{NB}}^T \bar{X}^{[+]} > \boldsymbol{d}_{\text{NB}}^T \bar{X}^{[-]}.$$

The NB classification rule τ_{NB}, constructed from $\mathbb{X}$ and associated class labels, is of the general form of Equation 4.1 and assigns the class label to a $d \times 1$ observation $\boldsymbol{x}$

$$\tau_{\text{NB}}(\boldsymbol{x}) = \begin{cases} +1 & \text{if} \quad \boldsymbol{d}_{\text{NB}}^T \boldsymbol{x} + \boldsymbol{d}_{\text{NB}}^T \left(\bar{X}^{[+]} - \bar{X}^{[-]} \right) > 0 \\ -1 & \text{if} \quad \boldsymbol{d}_{\text{NB}}^T \boldsymbol{x} + \boldsymbol{d}_{\text{NB}}^T \left(\bar{X}^{[+]} - \bar{X}^{[-]} \right) < 0 \end{cases} . \tag{4.8}$$

Because $\operatorname{diag} \hat{W}$ is a diagonal matrix, it is invertible so long as there are no variables in our dataset $\mathbb{X}$ that are constant in all $\mathbb{X}^{[\nu]}$ (zero within-class variance). The fact that $\operatorname{diag} \hat{W}$ is always invertible means that NB classification can be applied to cases with $n - \kappa < d$, where LDA cannot be used. Similarly to LDA, NB can be easily extended to more than two classes but we will only be considering the two-class case.

4.1.4 Distance Weighted Discrimination

Distance Weighted Discrimination (DWD) was introduced by Marron et al. (2007) as an approach to address 'data-piling' which may occur in Support Vector Machine (SVM) approaches. Data-piling occurs when multiple high-dimensional observations are projected to the exact same value, and is often a symptom of over-fitting effects in HDLSS data. As data-piling can be indicative of over-fitting effects, it is undesirable. We have presented linear classification in Equation 4.1 in terms of the projection of data into a direction vector $\boldsymbol{d}$. The motivation for DWD is based on an alternative perspective for thinking about linear classification rules in which we instead think about a hyperplane separating our classes in high-dimensional space, and consider the 'residuals' of the data to this hyperplane. These two perspectives of linear classification rules are equivalent. A hyperplane is a space of dimension one less than the dimension of the space within which it exists. If we consider $\boldsymbol{d}$ to be a normal vector to the hyperplane, or the hyperplane to be the space of all vectors orthogonal to $\boldsymbol{d}$ we can see that the projection of data onto the direction $\boldsymbol{d}$ are equivalent to the residuals of the same data from the hyperplane. To be precise, the residuals are the projection of the data onto the direction $\boldsymbol{d}$ plus a scalar β. The scalar β represents the location of the hyperplane on the line in the direction of $\boldsymbol{d}$. In HDLSS data it is often possible to linearly separate the two classes perfectly — i.e. to place a hyperplane such that all observations from one class are on one side of the hyperplane, and all observations from the other class are on the other side. In the case that the classes can be separated perfectly, a popular approach is to choose a hyperplane that maximises the minimum residual. This approach either only considers the smallest residual, or heavily weights the smallest residuals. Particularly in HDLSS data, this heavy weighting of the smallest residuals can cause multiple residuals to be exactly equal smallest, i.e. data-piling. DWD attempts to avoid the data-piling caused by this max-min approach by weighting larger residuals more heavily. DWD weights residuals based on their reciprocals — minimising the sum of reciprocal residuals, with an added penalty factor for observations on the wrong side of the hyperplane.

We now formulate DWD precisely, but first introduce some notation. If we denote the ith $d \times 1$ observation $\boldsymbol{x}_i$, and the ith class label y_i (with possible values of -1 and $+1$), we can define the ith residual

$$r_i^* = y_i(\boldsymbol{d}^T \boldsymbol{x}_i + \beta).$$

When the classes are perfectly split all the r_i^*s can be positive for particular choices of hyperplane, i.e. $\boldsymbol{d}$ and β. We define perturbed residuals

$$r_i = r_i^* + \epsilon_i$$

by adding positive error terms ϵ_i to the residuals — allowing the perturbed residuals to be positive even when the hyperplane does not split the classes perfectly, and thereby allowing us to pose the optimisation problem as in Equation 4.9. If we denote the vector of r_is $\boldsymbol{r}$ and the vector of ϵ_is $\boldsymbol{\epsilon}$, then given some penalty parameter C the DWD approach finds a solution to the optimisation problem:

$$\underset{\boldsymbol{d},\beta,\boldsymbol{r},\boldsymbol{\epsilon}}{\arg\min} \sum_i \left(\frac{1}{r_i} + C\epsilon_i\right) \tag{4.9}$$

under the conditions

$$||\boldsymbol{d}||^2 \leq 1, \quad r_i \geq 0, \quad \text{and} \quad \epsilon_i \geq 0 \quad \forall i.$$

Some comments:

- The condition that the vector $\boldsymbol{d}$ be a unit vector is relaxed to $|\boldsymbol{d}| \leq 1$ which makes the optimisation problem convex, but if the classes are perfectly separable the solution for $\boldsymbol{d}$ will be a unit vector for a sufficiently large penalty factor C.

- For $\boldsymbol{x}_i$ that lie on the correct side of the hyperplane, ϵ_i will be zero for a sufficiently large penalty factor C.

- We use the penalty factor recommended by Marron et al. (2007) — 100 divided by the median pairwise Euclidean distance between observations in one class to observations in the other class.

- As long as the penalty factor C is not too large, DWD will sometimes choose a hyperplane that does not perfectly split the observations, even in situations when it is possible to do so — something the max-min approach mentioned above will never do.

Let $\boldsymbol{d} = \boldsymbol{d}_{\text{DWD}}$ and $\beta = \beta_{\text{DWD}}$ be those values found to optimise the problem formulated in Equation 4.9. The DWD classification rule τ_{DWD}, constructed from $\mathbb{X}$ and associated class labels, is of the general form of Equation 4.1 and assigns the class label to a $d \times 1$ observation $\boldsymbol{x}$

$$\tau_{\text{DWD}}(\boldsymbol{x}) = \begin{cases} +1 & \text{if} \quad \boldsymbol{d}_{\text{DWD}}^T\boldsymbol{x} + \beta_{\text{DWD}} > 0 \\ -1 & \text{if} \quad \boldsymbol{d}_{\text{DWD}}^T\boldsymbol{x} + \beta_{\text{DWD}} < 0 \end{cases}. \tag{4.10}$$

The reciprocal weights used in the optimisation problem of Equation 4.9 cause DWD to take into account how well the two classes are separated overall, with less emphasis placed on the single smallest residual, in particularly when there are many similarly small residuals. Compared with the simple max-min approach mentioned above, this weighting used in Equation 4.9 results in DWD being less susceptible to over-fitting effects and data-piling, which can cause DWD to be particularly advantageous in HDLSS scenarios.

4.2 Preprocessing MALDI imaging data for Classification

In order to apply the methods introduced in Sections 4.1 and 4.3 to MALDI-MSI data, the data needs to be preprocessed and brought into a form such that these methods will be applicable. This involves two main points that require consideration:

- Variables — what are the measurements going to be. This point boils down to a question of how the m/z domain ought to be discretised.
- Observations — what are the objects to be classified. Spectra are perhaps not appropriate, as the objects for which real-world classification is of interest are the patients. So, a patient-wise representation of the data is needed.

We discuss our approach to each of these two points in Section 4.2.1 and Section 4.2.2 respectively.

In Chapter 2 and Chapter 3 we dealt primarily with the binary representation of MALDI-MSI data — sidestepping the problem of noise in other measures of peak presence (such as intensity, or SNR). Transforming the data to binary representation involves a significant loss of information, but we demonstrated that tissue types can still be effectively separated using the binary data despite this loss of information. It is possible, however, that this lost information could be of use in improving classification results. As our aim is to explore different approaches to the classification of MALDI-MSI data and to determine if any such approaches consistently perform better than others we will consider a variety of data types:

- Binary (presence/ absence of peaks),
- Intensity (peak height),
- Area (integrated peak volume),
- SNR, and
- Log-Intensity ($log(I + 1)$ where I is intensity).

The non-binary measures of peak-presence could potentially contain information important to the classification problem, and comparing classification performance on these different data-types ought to provide some insight into this.

4.2.1 Variables (Binning and Majority Rule)

We discussed the advantages and disadvantages of using data-dependent discretisation for constructing variables in Section 3.2. In the context of classification, it is appropriate to use data-independent discretisation (i.e. binning), as this allows for classification rules to be applied to new data unambiguously, as the same discretisation can be applied to any data. If data-dependent discretisation was used, how to apply a classification rule to new data would be somewhat ambiguous, and using data-independent discretisation avoids this problem.

However, as discussed in Section 2.6.2, binning can potentially remove important information in a small number of variables for which bin edges happen to fall in a region of high peak-density. In classification this small loss of information could potentially impact results, if the variables affected coincide with variables important to the classification problem in question. In order to address this potential loss

of important information, we suggest using several shifted bin locations, as explicitly defined in Algorithm A.3, applying any given classification approach to each shifted-bin dataset in parallel, and finally defining a 'meta classification rule' as the classification rule that assigns the class label agreed upon by the majority of the shifted-bin analyses. We will use three shifted-bin analyses as this will always guarantee a unique majority. We carry over the choice to use a bin size of $b = 0.25$ from the analysis of the ovarian cancer data and the discussion thereof in Section 2.3.1. All classification results we present in Chapter 5 are the result of using a 'meta classification rule' on three shifted-bin analyses resulting by bin location shifts of $-\frac{b}{3}$, 0, and $+\frac{b}{3}$ as per Algorithm A.3.

4.2.2 Observations (Averages and Cancer Annotation)

Until now all the data we have presented have had spectra as observations. In the endometrial data we wish to classify patients as LNM positive or negative, and so it is natural for observations to correspond to patients. The simplest way to construct such a 'patient-wise' representation is to average the spectra from each patient. We call this average spectra representation the 'patient-wise summarised data' or 'patient data', and exclusively use this representation of the data when considering any classification.

Note that so far we have predominantly been using the binary data, in which case these patient-wise averages produce within-patient proportions of occurrence. But, as mentioned in Section 4.2.1, for classification we will also consider using non-binary measures of peak-presence, and in these cases it is not obvious how to treat peak-absence, which is essentially a missing value problem. We present two options, and we will consider results of applying both in Chapter 5:

- Use the value zero to represent the absence of a peak.
- For each variable, average only present peaks (ignoring absence).

The two approaches above are not necessarily the best, but are the simplest. Note that all these analyses are on the basis of the peaklist data. Which peaks are 'present' and 'absent' is defined by the peak-picking algorithm, which uses a SNR threshold. The above two points could be interpreted in terms of this threshold rather than in terms of peak absence, e.g. 'the value of peaks with a SNR below the threshold are set to zero' or 'only peaks with a SNR above the threshold are averaged'.

One of the advantages of MSI data is the fact that spatial information that separates tissue types is preserved, as discussed in Chapters 2 and 3. Averaging spectra in order to produce a patient-wise representation of the data loses all information about within-patient tissue heterogeneity. It is natural to presume results could be improved by incorporating histological information separating tissue types, and thereby reducing variability in the patient data due to tissue heterogeneity. We take an approach similar to that of Mascini et al. (2015) — we have a pathologist annotate the tumour regions on an image of the H&E stained tissue, and restrict attention to spectra from tumour regions. We expect that restricting to the annotated spectra only should improve results as it should reduce the variability caused by tissue heterogeneity — comparing tumour tissue from one patient to tumour tissue of another patient should allow for differences between the patients to be detected more easily than including all the tissue from each patient. We present results of classification of the cancer annotated data in Section 5.3 and observe that although improvement is seen in some cases, in other cases restricting to the annotated spectra only can worsen results. The fact that restricting to cancer annotations does

not strictly improve classification results is surprising. One possible explanation for this surprising result can be observed from Table 1.5 in that there are a number of patients with very few cancer annotated spectra, so restricting to only cancer annotated spectra could artificially reduce our sample size, thereby explaining the increase in misclassification. An alternative explanation is that there could be useful information available in surrounding non-tumour tissues. Evidence supporting this intriguing possibility exists, for example Oppenheimer et al. (2010) have shown that histologically normal tissue adjacent to renal carcinoma tumours express many of the molecular characteristics of the tumour. This is a possibility that, as Oppenheimer et al. (2010) note, warrants further research as it could potentially relate directly to tumour recurrence post resection, which is a significant factor in patient survival.

4.3 Dimension Reduction

As mentioned in Section 4.1.2, classic classification methods such as LDA fail for HDLSS data. One approach to addressing this failure is to transform the data into a low-dimensional representation, typically a subspace, prior to classification. We will consider two methods for such dimension reduction:

- PCA, and
- Variable selection based on Canonical Correlation Analysis (CCA) ranking.

PCA is a commonly used variable reduction method, see Koch (2013, Chapter 2 and Section 13.3.2). PCA has been used for dimension reduction extensively, and specifically in the context of TMA MALDI-MSI data Mascini et al. (2015) have suggested PCA dimension reduction followed by LDA. CCA is an established method in multivariate statistics, see Koch (2013, Chapter 3). Koch and Naito (2010) have suggested the use of CCA for variable ranking, and we use an approach similar to that which they suggest. The slight deviation between our approach and that of Koch and Naito (2010) is that we centre the class labels, while Koch and Naito (2010) suggest using uncentred labels. We use an approach more similar to that described in Koch (2013, Section 13.3.1), but the difference between our approach to that of Koch and Naito (2010) is trivial in terms of practical results.

We introduce the known ideas of PCA and CCA in Section 4.3.1 and Section 4.3.2 respectively. We apply these variable reduction approaches to the endometrial data of Section 1.5.3 and consider the effect they have on the classification performance in Section 5.2.

4.3.1 PCA

Let S be the $d \times d$ sample covariance matrix of a $d \times n$ centred data matrix $\mathbb{X}$, and let S have rank r. As $\mathbb{X}$ is centred, $S = \frac{1}{n-1}\mathbb{X}\mathbb{X}^T$. Let the eigendecomposition, often called spectral decomposition, of S be

$$S = \Gamma\Lambda\Gamma^T, \tag{4.11}$$

where Γ is a $d \times r$ matrix whose columns are eigenvectors of S, and Λ is an $r \times r$ diagonal matrix of the eigenvalues of S $\lambda_1, \lambda_2, \ldots, \lambda_r$ such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$. Note that for the patient-data we consider for classification, as described in Section 4.2.2, the matrix S is singular as $n << d$. We call the columns of Γ

principal component directions, and we project the data into these directions in order to produce the PCA dimension reduced data of Definition 15.

Definition 15. k-dimension reduced PCA data: *Given a $d \times n$ centred data matrix $\mathbb{X}$, covariance matrix $S = \frac{1}{n-1}\mathbb{X}\mathbb{X}^T$ and the eigendecomposition notation of Equation 4.11, Let the $d \times k$ matrix Γ_k consist of the first k columns of Γ. Then for a given number of components k, the k-dimension reduced PCA data is*

$$\Gamma_k^T \mathbb{X}.$$

Note that Definition 15 is a function of the number of dimensions reduced to, k, and that each of the resulting dimensions correspond to linear combinations of the original variables. Each of these dimensions, or rows of the dimension reduced PCA data matrix, correspond to the projection of the original centred data into a principal component direction and are often called principal component scores. These new dimensions are not immediately interpretable as m/z values, as they correspond to combinations of many m/z values and do not have a direct interpretation in MS terms.

It is also quite common to produce the scaled data,

$$(\operatorname{diag} S)^{-\frac{1}{2}} \mathbb{X},$$

where $\operatorname{diag} S$ is a diagonal matrix whose diagonal is the same as that of S, and apply PCA to the scaled data. The covariance matrix of the scaled data is the correlation matrix of the original matrix. We do not consider PCA on the scaled data here as our focus is on comparing PCA to other methods, which are likely to have a bigger impact on the classification results than scaling would.

For a detailed discussion of the interpretation of PCA, including proofs for its theoretical properties, see Koch (2013, Chapter 2). In short, it can be shown that the k-dimension reduced PCA data maximises the variability retained from the original data. One of the problems with PCA as a dimension reduction method is that the variables in the transformed data do not have an obvious interpretation as they are linear combinations of the original variables. Furthermore, in a classification context the 'highest variance' directions are not necessarily the 'best' directions — they are not necessarily the dimensions most relevant to the classification problem in question. The CCA-based approach we describe in Section 4.3.2 is an example of an approach that attempts to leverage the extra information contained in the known class labels in order to find variables that are the most relevant to the classification problem. We would expect a method that makes use of the extra information contained in the class labels to result in better classification performance, and indeed the results in Chapter 5 reflect that the CCA-based approach results in better classification performance than PCA dimension reduction. PCA variable reduction is nonetheless a staple variable reduction method, and a useful baseline for comparison to other approaches.

4.3.2 CCA

As mentioned in Section 4.3.1, PCA is a purely variance based technique, using the variance of the data matrix $\mathbb{X}$ without using the class labels at all. In some contexts variance based variable reduction can be appropriate, for example, clustering. In classification however, there is no guarantee that the information differentiating

the classes will be contained in higher-variance components. Leek et al. (2010) discuss how exploratory analyses using PCA and other methods, including hierarchical clustering, can reveal that some of these high-variance components often represent unwanted variability such as batch effects — effects caused by differences in the environment during data acquisition, trace-contaminants in reagents, systematic operator errors, and other similar effects.

In contrast to PCA, the CCA-based variable ranking we propose in this section takes the class labels into account when ranking variables in order of importance. CCA is a general method for finding strong correlations between two subsets of variables. CCA is of particular interest when there is a natural partitioning of the variables by context. For example, Witten and Tibshirani (2009) demonstrate this principal on a dataset including measurements of both gene expression and DNA copy number for the same samples. Specifically, Witten and Tibshirani (2009) use an extension of CCA to find sparse linear combinations of these two sets of measurements that are highly correlated to each other. In our classification context we can consider the class labels as one set of variables, and the MALDI-MSI data as the other. In essence, the CCA-based variable ranking we propose ranks the variables of our data based on their correlation to the class labels.

In this section we introduce CCA in two parts. Firstly we introduce CCA in general. Secondly, we consider CCA in the specific context of two-class classification, and many of the general expressions simplify in this context.

In general

Let $\mathbb{X}_1$ and $\mathbb{X}_2$ be $d_1 \times n$ and $d_2 \times n$ data matrices respectively, corresponding to two sets of measurements or variables on the same n subjects or observations. For convenience, let $\mathbb{X}_1$ and $\mathbb{X}_2$ be centred — each row has mean zero. Then the sample covariance matrix of

$$\begin{bmatrix} \mathbb{X}_1 \\ \mathbb{X}_2 \end{bmatrix} \quad \text{is} \quad \frac{1}{n-1} \begin{bmatrix} \mathbb{X}_1 \\ \mathbb{X}_2 \end{bmatrix} \begin{bmatrix} \mathbb{X}_1 \\ \mathbb{X}_2 \end{bmatrix}^T = \frac{1}{n-1} \begin{bmatrix} \mathbb{X}_1\mathbb{X}_1^T & \mathbb{X}_1\mathbb{X}_2^T \\ \mathbb{X}_2\mathbb{X}_1^T & \mathbb{X}_2\mathbb{X}_2^T \end{bmatrix} = \begin{bmatrix} S_1 & S_{12} \\ S_{12}^T & S_2 \end{bmatrix}. \tag{4.12}$$

As mentioned in Section 4.3.1, we will often be dealing with singular covariance matrices in the classification setting. Let

$$S_1 = \Gamma_1 \Lambda_1 \Gamma_1^T \quad \text{and} \quad S_2 = \Gamma_2 \Lambda_2 \Gamma_2^T$$

be the eigendecompositions of S_1 and S_2 respectively, analogously to the eigendecomposition of S in Equation 4.11. We use the notation $S_1^{-\frac{1}{2}}$ and $S_2^{-\frac{1}{2}}$, and in order to avoid ambiguity in the cases when S_1 or S_2 are singular, we let

$$S_1^{-\frac{1}{2}} = \Gamma_1 \Lambda_1^{-\frac{1}{2}} \Gamma_1^T \quad \text{and} \quad S_2^{-\frac{1}{2}} = \Gamma_2 \Lambda_2^{-\frac{1}{2}} \Gamma_2^T. \tag{4.13}$$

The expressions for $S_1^{-\frac{1}{2}}$ and $S_2^{-\frac{1}{2}}$ in Equation 4.13 remain well defined in the case that either S_1 or S_2 is singular because of the way we introduced eigendecomposition in Equation 4.11 — that is if we let r_1 and r_2 be the ranks of S_1 and S_2 respectively, then Λ_1 and Λ_2 are $r_1 \times r_1$ and $r_2 \times r_2$ diagonal matrices of non-zero eigenvalues respectively, and similarly Γ_1 and Γ_2 are $d_1 \times r_1$ and $d_2 \times r_2$ matrices of eigenvectors. Allowing Λ_1 and Λ_2 to be of lower dimension, r_1 or r_2 instead of d, and allowing Γ_1 and Γ_2 to be non-square is what allows us to define $S_1^{-\frac{1}{2}}$ and $S_2^{-\frac{1}{2}}$ in such a way that they are still defined even when S_1 or S_2 is singular.

Given the expressions for $S_1^{-\frac{1}{2}}$ and $S_2^{-\frac{1}{2}}$ in Equation 4.13 we let the canonical correlation matrix

$$C = S_1^{-\frac{1}{2}} S_{12} S_2^{-\frac{1}{2}},$$

as in Koch (2013, Equation (3.13)). C is $d_1 \times d_2$. Let

$$C = P \Upsilon Q^T$$

be the singular value decomposition of C. For details on singular value decomposition see Koch (2013, Definition 1.12). If we let r be the rank of C then P, Υ, and Q are $d_1 \times r$, $r \times r$ and $d_2 \times r$ respectively. Let the diagonal entries of the diagonal matrix Υ be denoted $\upsilon_1 \geq \upsilon_2 \geq \ldots \geq \upsilon_r$, and the columns of P and Q $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_r$ and $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_r$ respectively.

CCA is related to what we will call the sphered data,

$$S_1^{-\frac{1}{2}} \mathbb{X}_1 \quad \text{and} \quad S_2^{-\frac{1}{2}} \mathbb{X}_2, \tag{4.14}$$

assuming that these inverses exist, and recalling that $\mathbb{X}_1$ and $\mathbb{X}_2$ denote the centred data. $\boldsymbol{p}_1$ and $\boldsymbol{q}_1$ are direction (unit) vectors such that the projection of the sphered datasets described in Equation 4.14 into these two directions respectively, i.e. $\boldsymbol{p}_1^T S_1^{-\frac{1}{2}} \mathbb{X}_1$ and $\boldsymbol{q}_1^T S_2^{-\frac{1}{2}} \mathbb{X}_2$, are maximally correlated to each other. The sphered data and the original data are related through the matrices $S_1^{-\frac{1}{2}}$ and $S_2^{-\frac{1}{2}}$, and these matrices can be used to transform the direction vectors $\boldsymbol{p}_k$ and $\boldsymbol{q}_k$ into the vectors

$$\boldsymbol{\phi}_k = S_1^{-\frac{1}{2}} \boldsymbol{p}_k \quad \text{and} \quad \boldsymbol{\psi}_k = S_2^{-\frac{1}{2}} \boldsymbol{q}_k \tag{4.15}$$

respectively. Projecting the sphered data of Equation 4.14 into the vectors $\boldsymbol{p}_k$ and $\boldsymbol{q}_k$ is equivalent to projecting the original data $\mathbb{X}_1$ and $\mathbb{X}_2$ into the vectors $\boldsymbol{\phi}_k$ and $\boldsymbol{\psi}_k$. As the $\boldsymbol{\phi}_k$s and $\boldsymbol{\psi}_k$s have this interpretation in terms of projecting the original data, they are commonly used in CCA rather than the $\boldsymbol{p}_k$s and $\boldsymbol{q}_k$s as introduced here and in Koch (2013, Chapter 3). It should be noted however that the $\boldsymbol{\phi}_k$s and $\boldsymbol{\psi}_k$s are not unit vectors as the $\boldsymbol{p}_k$s and $\boldsymbol{q}_k$s are. Also, the $\boldsymbol{\phi}_k$s and $\boldsymbol{\psi}_k$s do not have the interpretation as left and right eigenvectors of C as the $\boldsymbol{p}_k$s and $\boldsymbol{q}_k$s do. The absolute values of the entries of $\boldsymbol{\phi}_1$ and $\boldsymbol{\psi}_1$ can give us rankings of the variables of $\mathbb{X}_1$ and $\mathbb{X}_2$ respectively in order of their contributions towards the strongest correlation between the two sets of measurements or variables. We will use these CCA-based rankings for variable ranking prior to classification.

In two-class classification

There are a number of approaches to using the concepts from CCA as introduced above for variable ranking in a classification context. For example, Koch (2013, Section 13.3.1) suggest a variable ranking approach that directly applies the general approach we have introduced above to a regression context, while in Koch (2013, Section 13.3.3) a slightly modified approach is suggested for variable ranking in a classification approach. The approach we use is essentially that of Koch (2013, Section 13.3.3) but with centred class labels. There is some ambiguity concerning how to numerically code classes — which are categorical in nature but must be represented numerically in order to be used in the context of CCA. In the two-class case, Marron et al. (2007) uses numeric labels of -1 and $+1$ as this is convenient for the formulation of their optimisation problem discussed in Section 4.1.4. Koch (2013) and some references therein use labels of 0 and 1 as this can simplify the

formulation of methods such as LDA. Which of these two options is better is not obvious. One of the reasons we modify the approach suggested in Koch (2013, Section 13.3.3) and use centred class labels is that using centred class labels causes our variable ranking method which we describe here to be invariant to choice of class labels — making this ambiguity irrelevant in the two-class case. The deviation we make from the suggested approach of Koch (2013, Section 13.3.3) is minor, and we would not expect our approach to produce greatly different results to that of Koch (2013, Section 13.3.3).

Let $\mathbb{X}_1 = \mathbb{X}$ be our $d \times n$ (centred) data matrix, and $\mathbb{X}_2 = \mathbb{Y}$ be our $1 \times n$ vector of (centred) class labels. The two-class case is particularly simple, as we can let $\mathbb{Y}$ be a $1 \times n$ binary vector whose entries are either $-\frac{2n_+}{n}$ or $\frac{2n_-}{n}$, coding for the two classes (of sizes n_- and n_+ respectively). These values of $\mathbb{Y}$ correspond to using class labels of -1 and $+1$, and then centring. Following the same process as in the general case above, the sample covariance matrix of

$$\begin{bmatrix} \mathbb{X} \\ \mathbb{Y} \end{bmatrix} \quad \text{is} \quad \frac{1}{n-1} \begin{bmatrix} \mathbb{X}\mathbb{X}^T & \mathbb{X}\mathbb{Y}^T \\ \mathbb{Y}\mathbb{X}^T & \mathbb{Y}\mathbb{Y}^T \end{bmatrix} = \begin{bmatrix} S_X & S_{XY} \\ S_{XY}^T & S_Y \end{bmatrix}.$$

Note that these terms are simplified in this context — S_{XY} is a $d \times 1$ vector, and S_Y is the scalar $\frac{4n_-n_+}{n(n-1)}$ corresponding to the variance of the entries of $\mathbb{Y}$. Using this notation, let

$$S_X = \Gamma\Lambda\Gamma^T$$

be the eigendecomposition of S_X. The canonical correlation matrix is found in the same way,

$$C = S_X^{-\frac{1}{2}} S_{XY} S_Y^{-\frac{1}{2}},$$

but is now a $d \times 1$ vector, so $\boldsymbol{p}_1$ is simply C normalised to length one and $\boldsymbol{\phi}_1 = S_X^{-\frac{1}{2}} \boldsymbol{p}_1$ as in Equation 4.15. We use the absolute values of the entries of $\boldsymbol{\phi}_1$ to construct a ranking of the variables of $\mathbb{X}$, and select the highly ranked variables to make the k-variable selected CCA data of Definition 16. Note that

$$\boldsymbol{\phi}_1 = S_X^{-\frac{1}{2}} \boldsymbol{p}_1 = S_X^{-\frac{1}{2}} \left(\frac{1}{|C|} C \right) = \frac{1}{|C|} S_X^{-\frac{1}{2}} \left(S_X^{-\frac{1}{2}} S_{XY} S_Y^{-\frac{1}{2}} \right) \propto S_X^{-1} S_{XY}. \tag{4.16}$$

Because we only use $\boldsymbol{\phi_1}$ for ranking variables, the ranking is invariant to multiplying $\boldsymbol{\phi_1}$ by a non-zero constant, and so we can use the simplified expression in Equation 4.16 for our ranking.

Definition 16. k-variable selected CCA data: *Given a $d \times n$ centred data matrix $\mathbb{X}$, a $1 \times n$ vector of centred class labels $\mathbb{Y}$ and covariance matrices $S_X = \frac{1}{n-1}\mathbb{X}\mathbb{X}^T$ and $S_{XY} = \frac{1}{n-1}\mathbb{X}\mathbb{Y}^T$, calculate the $d \times 1$ ranking vector $\boldsymbol{\phi} = S_X^{-1} S_{XY}$. If S_X is singular, let S_X^{-1} be the Moore-Penrose psuedoinverse. For details on the Moore-Penrose psuedoinverse, see Penrose (1955); Ben-Israel and Greville (2003). Let $\boldsymbol{\mathfrak{d}}$ be a $d \times 1$ vector containing ones in positions corresponding to the k elements of $\boldsymbol{\phi}$ with highest absolute values and zeros elsewhere. The k-variable selected CCA data is the submatrix $T_{\boldsymbol{\mathfrak{d}}}^T \mathbb{X}$.*

See Definition 7 for details on the submatrix notation used in Definition 16. In comparison to the PCA dimension reduction approach of Section 4.3.1, the CCA-based variable selection approach of Definition 16 results in data whose variables can be directly interpreted as corresponding to m/z values, rather than linear combinations of m/z values. Also, we expect the CCA-based approach to dimension reduction to produce better classification results than the PCA approach of Section 4.3.1 because CCA uses the information we have from the class labels. The results we present in Chapter 5 confirm this expectation.

4.4 Normalisation

In Chapter 2 we motivate the use of the binary data by talking about the noise that is typically inherent in non-binary measures of peak-presence. However, as discussed in Section 4.2, we will consider non-binary measures of peak-presence as well as the binary data. As such in this section we introduce a new approach that attempts to 'normalise' these non-binary measures of peak-presence using the internal calibrants of Gustafsson et al. (2012) in an attempt to reduce unwanted variability. We will refer to intensity, but the same approach could equally apply to any other measure of peak presence (SNR, or integrated area for example).

Both peptide applications we consider, ovarian and endometrial cancer, have internal calibrants added in the sample preparation step. These internal calibrants are used to calibrate the m/z measurements, as described by Gustafsson et al. (2012). The internal calibrants are sprayed onto the tissue evenly during sample preparation and so we know that there should be the same concentration of each calibrant at any given location on the tissue. Given the calibrant concentrations should be constant, we would expect the corresponding intensities to be constant. We assume that the overall intensity measurements over an entire spectrum are all affected to the same degree by extraneous variables such as matrix crystallisation and total signal suppression. We can use the calibrant intensities to estimate this systematic effect for each spectrum and adjust for these effects. We call this adjustment 'normalisation'.

4.4.1 The Model

The underlying model for our data is

$$x_{ij} = \mu_{ij} s_j \epsilon_{ij} \tag{4.17}$$

for the observed intensity, x_{ij}, of variable or molecular species (m/z) i in spectrum j, where μ_{ij} is an intensity representative of the true concentration of the species i present in spectrum j, s_j is the systematic error we would like to estimate and compensate for, and ϵ_{ij} is random noise, which we will assume to be log-normal, as is fairly typical in such data. The fact that s_j is independent of variable i reflects our assumption that all intensity measurements in any given intensity should be affected to the same degree. Similarly, if we let $\mathcal{D}$ denote the set of variables that correspond to the internal calibrants, then for $i \in \mathcal{D}$ we can represent our assumption that calibrants are evenly distributed across the tissue by omitting the spectra dependence, i.e. $\mu_{ij} = \mu_i$. These two assumptions are what allow us to simplify the model sufficiently such that it is no longer over-parametrised and can now actually be fit to data in order to estimate the parameters we are interested in, i.e. the s_j.

We consider the log-model for the calibrants,

$$\log(x_{ij}) = \log \mu_i + \log s_j + \log \epsilon_{ij} \qquad i \in \mathcal{D}. \tag{4.18}$$

Let us have n spectra, d calibrants, and let us assume each spectra contains all d calibrants. This log-model can be written using matrices as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.19}$$

where the $dn \times (n + d - 1)$ design matrix $\boldsymbol{X}$ is defined in Equation 4.20, and the $(n + d - 1) \times 1$ parameter vector $\boldsymbol{\beta}$, and the $dn \times 1$ response vector $\boldsymbol{y}$ are defined in

Equation 4.21.

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & & 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & & \ddots & \vdots & \vdots & & & & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & & 0 & 1 & 0 & 0 & \dots & 0 \\ \vdots & & \ddots & \vdots & \vdots & & & & \vdots \\ 0 & 0 & \dots & 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & & 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & & \ddots & \vdots & \vdots & & & & \vdots \\ 0 & 0 & \dots & 1 & 0 & 1 & 0 & \dots & 0 \\ \vdots & & & \vdots & \vdots & & \ddots & & \vdots \\ 1 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1 \\ 0 & 1 & & 0 & 0 & 0 & 0 & \dots & 1 \\ \vdots & & \ddots & \vdots & \vdots & & & & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & 0 & \dots & 1 \end{bmatrix}, \tag{4.20}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \log(\mu_1) + \log(s_1) \\ \log(\mu_2) + \log(s_1) \\ \vdots \\ \log(\mu_d) + \log(s_1) \\ \log(s_2) - \log(s_1) \\ \log(s_3) - \log(s_1) \\ \vdots \\ \log(s_n) - \log(s_1) \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} \log(x_{11}) \\ \log(x_{21}) \\ \vdots \\ \log(x_{d1}) \\ \log(x_{12}) \\ \log(x_{22}) \\ \vdots \\ \log(x_{d2}) \\ \vdots \\ \vdots \\ \log(x_{1n}) \\ \log(x_{2n}) \\ \vdots \\ \log(x_{dn}) \end{bmatrix}. \tag{4.21}$$

$\boldsymbol{\epsilon}$ of Equation 4.19 is assumed to be a vector of independent identically distributed normal variables (white noise). The formulation of Equation 4.19 is the standard form for a linear regression model with two categorical independent variables — in this case, the two categorical variables essentially correspond to spectrum and calibrant. The parametrisation defined by our choice of $\boldsymbol{\beta}$ in Equation 4.21 is not unique, but the estimation it yields for the s_j is unique. The rows of $\boldsymbol{X}$ and y as in Equations 4.20 and 4.21 correspond to the individual intensity measurements and are split into n blocks of d, each block corresponding to a spectrum, and within each block each row corresponding to a calibrant. The precise form of $\boldsymbol{X}$ is determined by the parametrisation chosen for $\boldsymbol{\beta}$ and the model assumptions discussed above. For more details on regression see Casella and Berger (2002). In this context, it is sufficient to understand that this formulation allows for the parameters of interest to be estimated within a well established statistical paradigm. Specifically, from linear regression we know that the least squares estimate for the parameter vector

$\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{4.22}$$

Most linear regression implementations involve numeric computation of the matrix inverse $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, which here is a $(n+d-1)\times(n+d-1)$ matrix. In our case $n > 10,000$, which leads to very slow computation. This matrix is of a very particular block-matrix form, for which we have derived an analytic form for its inverse — see Appendix C for the derivation. Having an analytic form for this inverse allows for much faster computation, as numeric estimation is not necessary, and this speeds up computations by several orders of magnitude. Note that for this design matrix $\boldsymbol{X}$,

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} n & 0 & \dots & 0 & 1 & 1 & \dots & 1 \\ 0 & n & & 0 & 1 & 1 & \dots & 1 \\ \vdots & & \ddots & \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & n & 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 & d & 0 & \dots & 0 \\ 1 & 1 & & 1 & 0 & d & \dots & 0 \\ \vdots & & \ddots & \vdots & \vdots & & \ddots & \vdots \\ 1 & 1 & \dots & 1 & 0 & 0 & \dots & d \end{bmatrix},$$

is of the form described in Equation C.1 — specifically $\boldsymbol{X}^T\boldsymbol{X} = A(n, d, d, n-1)$. The general result from Appendix C gives us the relatively simple form

$$\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right]_{ij} = \begin{cases} \frac{d+n-1}{dn} & i,j \in [1,d] & i = j \\ \frac{2}{d} & i,j \in [d+1, d+n] & i = j \\ \frac{n-1}{dn} & i,j \in [1,d] & i \neq j \\ \frac{1}{d} & i,j \in [d+1, d+n] & i \neq j \\ \frac{-1}{d} & \begin{matrix} i \in [1,d],\ j \in [d+1,d+n] \\ j \in [1,d],\ i \in [d+1,d+n] \end{matrix} & \text{or} \end{cases}$$

This relatively simple form for $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ allows us to derive $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ by matrix multiplication and from this we can use Equation 4.22 to find analytic solutions to the least squares estimates of the parameters:

$$\widehat{\log(\mu_1) + \log}(s_1) = \bar{x}_i + \frac{1}{d}\sum_{k=1}^{d}(\log(x_{k1}) - \bar{x}_k)$$

and

$$\widehat{\log(s_j) + \log}(s_1) = \frac{1}{d}\sum_{i=1}^{d}(\log(x_{ij}) - \bar{x}_i) - \frac{1}{d}\sum_{i=1}^{d}(\log(x_{i1}) - \bar{x}_i),$$

where

$$\bar{x}_i = \sum_{j=1}^{n}\log(x_{ij})$$

These analytic solutions to the least squares estimates agree with the intuitive estimates

$$\widehat{\log(\mu_i)} = \bar{x}_i \quad \text{and} \quad \widehat{\log(s_j)} = \frac{1}{d}\sum_{i=1}^{d}(\log(x_{ij}) - \bar{x}_i). \tag{4.23}$$

Ultimately all we are interested in is estimating the s_j, which represent the unwanted variability in our data we wish to adjust for. The expression for $\widehat{\log(s_j)}$ as in Equation 4.23 provides us with a way to estimate these s_j as $\hat{s}_j = e^{\widehat{\log(s_j)}}$, and all the preceding work in this section was simply about justifying the choice to estimate the s_j in this way. From here onwards, all that we are interested in is the fact that we can estimate s_j using the intensity measurements of the calibrants $\mathcal{D}$, and we can adjust for unwanted variability in the data by replacing the intensity measurements in our data with the normalised intensities

$$x^*_{ij} = \frac{x_{ij}}{\hat{s_j}}. \tag{4.24}$$

4.4.2 Proof of Principle on the motivating dataset A3

In Section 4.4.1 we established a model that we can use to adjust for unwanted variability by using the intensity measurements of our calibrants $\mathcal{D}$ to obtain the normalised intensities (Equation 4.24). Now we are interested in applying this to a motivating dataset to validate that the method actually reduced the unwanted variability.

A natural way to test the performance of this normalisation would be to look at a m/z corresponding to a peptide that is uniformly distributed across the tissue before and after normalisation and (hope to) observe a reduction in the variability or spread of the intensity values observed for that m/z value. However, the only peptides expected to be uniformly distributed across the tissue are the internal calibrants. As the internal calibrants are used to estimate the parameters in the normalisation model, looking at the internal calibrants before/ after normalisation would give an optimistic measure of the effectiveness of the normalisation. We use an approach similar to that of LOO CV as described in Section 4.1.1 where for each of the four calibrants we fit the normalisation model using the other three and normalise the intensities of the calibrant we left out of the model-fit step. We can then consider the intensities of the internal calibrants before and after normalisation and thereby estimate the effectiveness of the normalisation. This estimation should infact be conservative, as in each case the model-fit is done on the basis of three calibrants. When the normalisation is done for the whole data, the model-fit step will be performed using all four calibrants, which should give better estimates than using only three. Figure 4.2 demonstrates the expected trend in these results — normalisation causing a reduction in the variability and range of intensity values for each calibrant. Although encouraging, the reduction in variability shown in Figure 4.2 is small relative to the total variability. We show the effect the normalisation has on classification results in Section 5.3. Normalisation improves classification results in some cases, but in other cases it worsens results. Overall, there does not seem to be a consistent trend showing that normalisation has an effect on classification results, although it is possible that our sample-size is simply too small to detect such a trend.

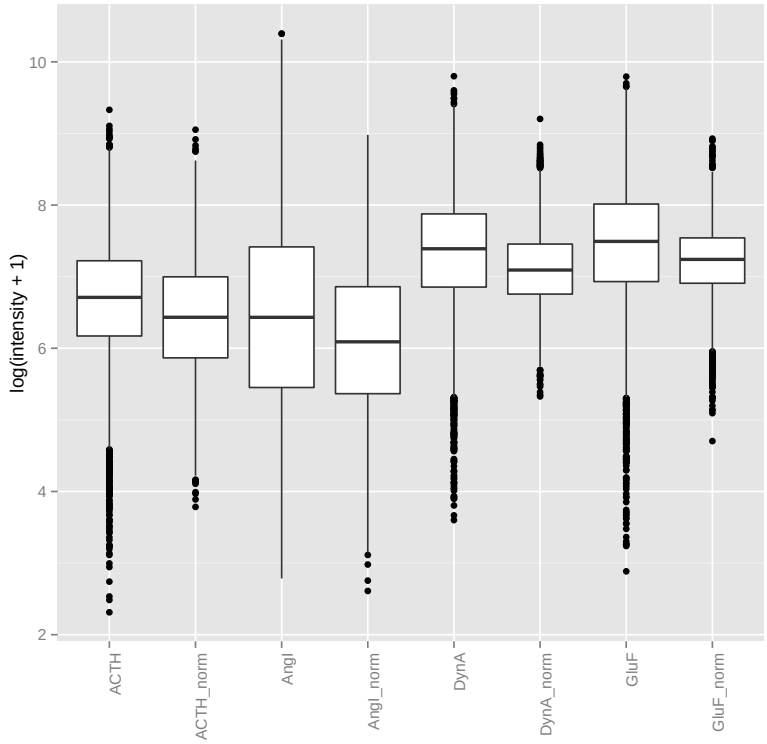


Figure 4.2: Boxplots of log-intensity on the y-axis is plotted for the four calibrants before and after normalisation. The x-axis separates between the calibrants, alternating between before-normalisation and after-normalisation results. For each calibrant, the normalisation model is fit using the other three calibrants, so as to avoid over-fitting effects. Spectra are restricted to only those including peaks for all 4 calibrants.

4.5 Summary

In this chapter we introduced and discussed a number of different points relating to the classification of MALDI-MSI data. In Chapter 5 we will apply these ideas to the TMA data of Section 1.5.3, evaluating and comparing many different classification schemes, but first here we will summarise the points covered in this chapter as they relate to the results in Chapter 5. In Section 4.1 we introduced CV and three classification methods: LDA, NB and DWD. We will exclusively use LOO CV for evaluating the performance of the various classification schemes we consider, and in each case we will compare the performance of the three classification methods introduced.

In Section 4.2 we introduced our approach to preprocessing MALDI-MSI data prior to classification, which involves two main discussion points: discretisation for constructing variables, and averaging of observations. We will use binning to discretise the m/z domain and construct variables, and we will replicate each classification three times, each differing only by a shift in bin locations. In each case we ultimately use the classification rule resulting from taking the majority result of the three parallel shifted-bin classifications. When averaging spectra, we include zeroes for absent peaks. As mentioned in Section 4.2, there are a number of data types that could be used, and we will compare results using five different data types, binary and four different non-binary data types (intensity, area, SNR, and log-intensity). Furthermore, for the non-binary data types we will also consider an alternative averaging scheme in which absent peaks are omitted rather than included as zeros, and compare results between these averaging schemes.

In Section 4.4 we introduced an approach to normalisation that could reduce unwanted variability in the non-binary data types, and in Chapter 5 we will consider results with and without such normalisation. For the binary data type, we will also consider the spatial smoothing introduced earlier in Section 2.5, to similar effect.

Finally, and perhaps most crucially, MALDI-MSI data are high-dimensional, and this poses a challenge for classification. In Section 4.3 we introduced two approaches to dimension reduction: a PCA based projection approach, and a CCA-based variable ranking approach. We will explore the results of applying these approaches in Chapter 5. So, in summary, in Chapter 5 we will consider each combination of the following options:

- **Dimension reduction approach** (PCA, CCA, or no dimension reduction) as introduced in Section 4.3,
- **Classification method** (NB, LDA, or DWD), *note: LDA cannot be used if no dimension reduction is performed, as discussed in Section 4.1.*
- **Spectra included in patient-averages** (all, or only annotated tumour spectra) as discussed in Section 4.2.2,
- **Data type** (area, binary, intensity, log-intensity, or SNR),
  - When non-binary data types are used,
    - **Normalisation** (with, or without) as described in Section 4.4, and
    - **Treatment of absent peaks when averaging** (include as zeros, or ignore) as discussed in Section 4.2.2,
  - When binary data is used, **Spatial smoothing** ($\tau = 0$, $\tau = 0.15$, or $\tau = 0.25$) as described in Section 2.5,

Chapter 5

Classification of Lymph Node Metastasis in Endometrial Cancer

In Chapter 4 we introduced a variety of classification methods, as well as discussing preprocessing and variable reduction of MALDI-MSI data prior to classification. In this chapter we will consider the application of these classification methods to the TMA data of Section 1.5.3 and in this context we will discuss the effects various preprocessing decisions have on classification performance. After presenting and discussing some initial results using each of the data types we will consider (binary, intensity, log-intensity, SNR, area) in Section 5.1, we consider variations in preprocessing, including dimension reduction approaches, which could potentially improve the classification performance. In Section 5.2 we apply the dimension reduction approaches described in Section 4.3. In Section 5.3 we consider variations in other preprocessing options prior to classification, including:

- Making use of histopathological annotations to restrict spectra that are included in patient-averages to only spectra from cancerous tumour tissue, removing spectra from non-tumour tissue.
- For the non-binary data we consider:
  - Normalisation using the internal calibrants, as described in Section 4.4, and
  - An alternative method for treating absent peaks when averaging spectra from each patient.
- For the binary data we consider the spatial smoothing of Section 2.5.

After having considered overall trends in the effects these preprocessing options have on classification performance, we take a closer look at some of the results that achieved the best LOO-misclassification in Section 5.4, and discuss how use of the CCA-based variable ranking approach to dimension reduction has the added benefit of identifying variables (m/z values) important for the classification — potential targets for follow up studies. In Section 5.5 we introduce a heuristic that lends some insight into the stability of linear classification rules, analogous to the concept of leverage from linear regression.

In Sections 5.1-5.5 we present many results, but generally make only a few sparse conclusions — the primary goal in these sections is to present the results. Having presented the results, we discuss overall trends and conclusions in Section 5.6. Ultimately our main conclusion is that the most important factor in determining classification performance is the choice of dimension reduction approach, and that

CCA-based variable selection performs very well. A couple of other factors, particularly classification method and data type also seem to have strong effects on classification performance, but with strong interaction effects that make predicting which option will perform well in any given situation somewhat difficult. Ultimately, our recommendation is to try several different variations to see which perform the best, but in our data CCA-LDA on the log-intensity data seems to perform very well.

The results presented in Sections 5.1-5.5 use the endometrial data described in Section 1.5.3 to explore the trends mentioned above, but it is of interest to explore if these trends generalise to the analysis of MALDI-MSI TMA data, or if they are specific to the endometrial data. To this end, we reproduced all analyses on the vulvar data, also introduced in Section 1.5.3, and these results are included in Appendix D. Some of the lesser trends are contradicted in the vulvar data. However, the main conclusions are all supported by the vulvar data results: that dimension reduction plays the biggest role in determining classification performance, that CCA-based variable selection performs very well, and that CCA-LDA on the log-intensity data consistently performs very well.

5.1 Data Processing and Initial Results

As discussed in Section 1.5.3, the endometrial cancer data consists of four datasets total: two sections (technical replicates) of two TMAs were analysed. In this analysis, we consider all four of these datasets together — making no distinction between them. If there exist batch effects between these datasets, this could be problematic, but we believe the methodology is sufficiently reproducible that any batch effects between datasets should be negligible. Furthermore, we are attempting to demonstrate that MALDI-MSI methodology can be used to predict patients LNM status, and if the methodology produces large batch effects such prediction would likely be impossible regardless. As discussed in Section 1.5.3, after consideration of the patient clinical data it was determined that 43 patients suitable for the study are represented across the two TMAs. Of these 43 patients, 16 are lymph node metastasis positive, 27 are negative. Details on the endometrial cancer project from which these data originate are available in Mittal et al. (2016).

As discussed in Section 2.1 we bin peaks with a bin width of 0.25 Da and, as discussed in Section 4.2.1, all analyses are replicated in parallel using bin locations shifted by $-\frac{0.25}{3}$ Da and $+\frac{0.25}{3}$ Da as in Algorithm A.3 to compensate for the fact that the binning is data-independent. We discussed the reasons why using multiple shifted-bin analyses is important in Section 2.6.2, and details are included in Section A.5. The bins represent variables in these data. Initially, for each patient, all spectra were averaged for each m/z bin. These averages are assembled into data matrices with $n = 43$ columns corresponding to the patients represented in the study, and $d = 4582$ rows corresponding to non-empty m/z bins ($d = 4570$ and $d = 4584$ in the two shifted-bin analyses respectively). These matrices are HDLSS ($n < d$) and so, as discussed in Section 4.1, LDA cannot be applied. NB and DWD can, however, and the LOO misclassification (as discussed in Section 4.1.1) of applying these two classification methods to these data are shown in Figure 5.1 for each of the different data types mentioned in Section 4.2. As we are performing three shifted-bin analyses in parallel, each result reported is the result of a majority 'meta-classification' rule combining the three classification results obtained from each of the parallel analyses.

Figure 5.1 shows that DWD strictly outperforms NB in this HDLSS context.

Figure 5.1: **Classification Without Dimension Reduction.** LOO misclassification on the y-axis using DWD or NB vs. data type on the x-axis.

DWD was specifically developed with the shortcomings of classical methods such as NB in exactly such high-dimensional cases in mind, so the fact that DWD outperforms NB here is perhaps unsurprising. The other interesting feature of Figure 5.1 is that the area and intensity data seem to outperform the other data types.

5.2 Dimension Reduction

In Section 4.3 we introduced two methods for dimension reduction: PCA, where the centred data are projected into the first k principal component directions (maximising variance) and CCA variable ranking, where variables in the data are ranked according to their correlation to the class labels (LNM status in this case) and the first k 'most important' variables are selected. Figure 5.2 shows the application of these two dimension reduction methods to the endometrial data (representing results that use PCA and CCA with solid and dashed lines respectively), prior to classification by the three methods introduced in Section 4.1: LDA, NB, and DWD. Figure 5.2 identifies results produced using a given data type with a single colour, and shows results for a range of values for the number of dimensions to reduce too, k, from 1 to 45. In the PCA dimension reduced data, represented by dashed lines in Figure 5.2, k is the number of principal components as discussed in Section 4.3.1. In the CCA variable selected data, represented by solid lines in Figure 5.2, k is the number of variables selected, as discussed in Section 4.3.2.

Note that LDA results exist for only $k \leq 40$, this is not only the maximum k such that $\hat{W}$ is invertible in practice, but infact the theoretical maximum for performing LOO CV. In general, $\hat{W}$ is singular for $n - \kappa < d$, but when using LOO CV n is replaced with $n - 1$ as in each case one observation is 'left-out' so, for LOO CV, $\hat{W}$ is singular when $d > n - \kappa - 1 = 43 - 2 - 1 = 40$. Similarly note that the PCA

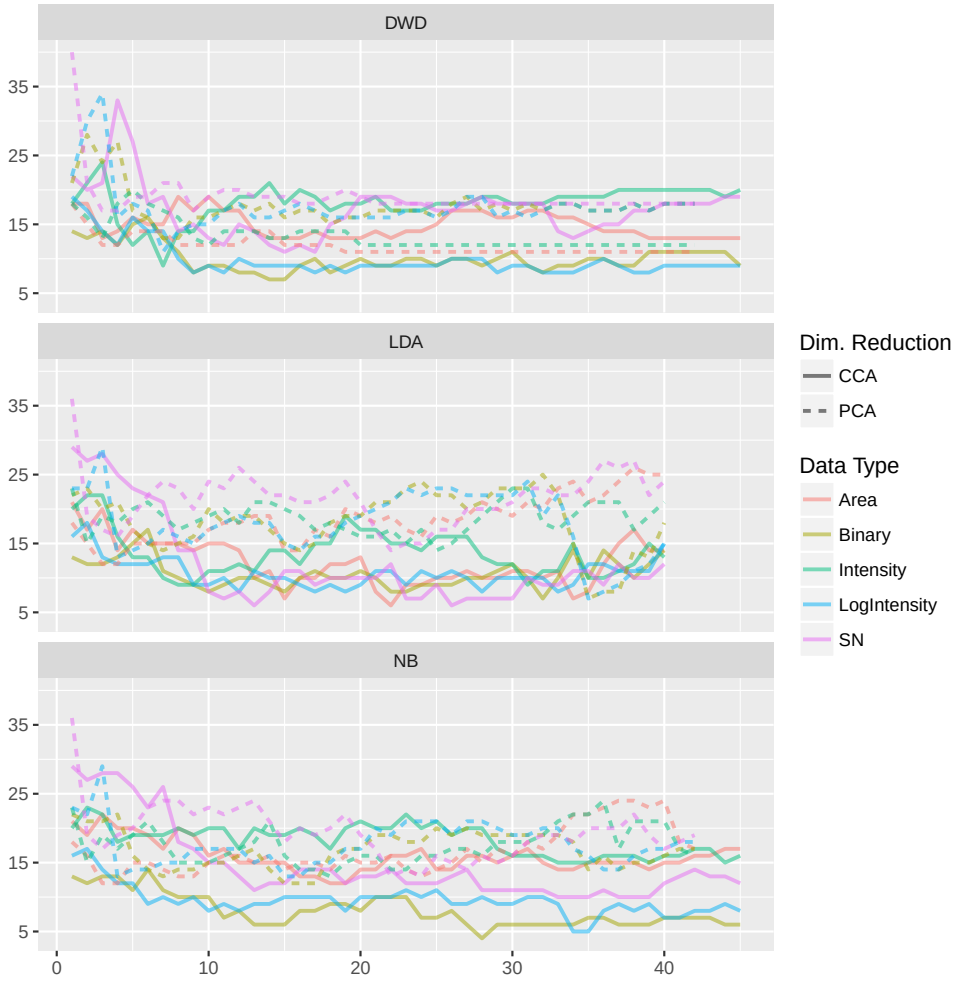


Figure 5.2: **Classification of Dimension Reduced Data.** LOO misclassification on the y-axis vs. the number of principal components on the x-axis for the PCA dimension reduced data, or the number of variables retained for the CCA variable reduced data. The results from using each classification method (NB, LDA, and DWD) are shown in seperate panels. Within each panel, results from using each data type are identified by colour. The LOO misclassification refers to the number of incorrectly classified patients out of 43.

dimension reduced results exist for $k \leq 42$, as the PCA dimension reduced data with $k = 42$ includes all principal components and 100% of the variance in the original data is preserved.

DWD shows less variation as k increases in its LOO misclassification compared with the other two classification methods, having quite stable LOO misclassification for $k \geq 20$. For the PCA dimension reduced data (dashed lines) in particular, the LOO misclassification stabilises to a single value in several cases, specifically:

- the PCA dimension reduced area data (dashed red line) achieves a stable minimum LOO misclassification by DWD of 11 for $k \geq 19$,
- and similarly the PCA dimension reduced intensity data (dashed green line) first achieves its minimum LOO misclassification by DWD of 12 at $k = 10$, and stabilises for $k \geq 20$.

This stable behaviour suggests that it suffices to consider the parsimonious models with $k = 19$ and $k = 20$ respectively in these cases.

In contrast to these stable minima, the PCA dimension reduced binary (yellow), log-intensity (blue), and SN (purple) data achieve minimum values of LOO misclassification by DWD at less than 10 components, specifically:

- the PCA dimension reduced binary data (dashed yellow line) achieves its minimum LOO misclassification by DWD of 13 at $k = 7, 8$,
- the PCA dimension reduced log-intensity data (dashed blue line) achieves its local minimum LOO misclassification by DWD of 11 at $k = 7$, and
- the PCA dimension reduced SN data (dashed purple line) achieves its minimum LOO misclassification by DWD of 17 at $k = 3, 4, 9$, and stabilises at a higher LOO misclassification of 18 for $k \geq 21$.

In contrast to the DWD results, NB and LDA exhibit relatively more instability as the number of components, k, increases. Of particular note is the minimum LOO misclassification of 7 achieved by LDA on the PCA dimension reduced data, with $k = 35$ for both the binary and log-intensity data (dashed yellow and blue lines respectively). This LOO misclassification of 7 is the best result achieved using PCA dimension reduction of all the results shown in Figure 5.2.

The broad trend for CCA variable reduced results is for the LOO misclassification to improve as k is increased until around $k = 10$, at which point a somewhat unstable minimum is achieved. Overall, CCA variable selection seems to outperform PCA dimension reduction, although in some cases this is not entirely clear. CCA outperforming PCA can be clearly observed in the LDA results of Figure 5.2, where the CCA results almost strictly outperform the PCA results. CCA variable selection also achieves minimum LOO misclassifications less than 7, achieving a minimum LOO misclassification of 6 using LDA with the area and SN data, and a minimum LOO misclassification of 4 using NB with the binary data.

Figure 5.2 contains an enormous amount of information and we are often only interested in the 'optimal result' that is the lowest LOO misclassification. From here on, we will display dimension-reduced results only for the optimal choice for the number of dimensions, k, in each case — the k that achieves the best LOO misclassification. In cases when there are multiple k that achieve equal best LOO misclassification, we choose the smallest — the most parsimonious. This will allow us to visualise the results we are interested in, while not overly crowding figures. Using these 'optimal results' will be useful in visualising the effect of various preprocessing options we will consider.

5.3 Varying Preprocessing Parameters

In this section, we will consider results achieved using alternative preprocessing alternatives, prior to dimension reduction and classification. First we will consider restricting to just spectra annotated as cancer when averaging spectra for each patient, as this is a preprocessing option for all the data types we consider. Next we consider some additional preprocessing options which are specific to either the binary, or non-binary data. For the binary data we consider spatial smoothing as described in Section 2.5. For the non-binary data types we consider normalisation, as described in Section 4.4, and alternative treatment of missing values when averaging.

5.3.1 Cancer Annotation

The first preprocessing option we consider, which applies in all data types, is whether to restrict attention to the spectra from regions of tissue annotated as tumour by the pathologist, or to include all spectra when averaging spectra from each patient. All results shown in Figures 5.1 and 5.2 include all spectra when averaging, regardless of annotation. To recap, Figure 5.1 shows 10 results without any dimension reduction corresponding to five different data types classified using NB and DWD. Figure 5.2 shows 30 cases corresponding to every combination of data type, dimension reduction approach (PCA or CCA) and classification method (LDA, NB, or DWD). As discussed at the end of Section 5.2, each of these 30 cases has an optimal choice for the number of dimensions k. The 10 cases with no dimension reduction combined with the 30 optimal cases corresponding to the cases represented in Figure 5.2 constitute the 40 points connected with solid lines in Figure 5.3. For each of these 40 cases, Figure 5.3 also shows the results when restricting to only annotated tumour spectra, and these alternative results are identified by being connected by dashed lines.

From Figure 5.3 we can see that when using CCA variable selection, restricting to only spectra from annotated tumour tissue seems to improve results — the notable exception to this being when NB is used on either the binary and log-intensity data (yellow and blue lines). When no dimension reduction step is included restricting to only annotated tumour spectra seems to have no noticeable trend in its effect, but when PCA dimension reduction is used restricting to only annotated tumour spectra seems to worsen results — the exceptions being the intensity and area data (green and red lines).

Another notable trend in Figure 5.3 is that LDA seems to perform better than NB or DWD on the CCA variable reduced data, with the the same two exceptions noted above — NB performs better on both the binary and log-intensity data, including all spectra (solid yellow and blue lines). Koch (2013, Section 13.3.3) contains a discussion of some intuition that may be of interest with regards to the link between LDA and CCA. It is also interesting that this trend of LDA performing better than NB and DWD does not apply to the PCA dimension reduced data however, with the same exceptions — the binary and log-intensity data including all spectra (solid yellow and blue lines).

Overall there seems to be no conclusive trend in the effect restricting to cancer annotated spectra only has on classification performance. However although not showing an overall effect this variation does show interesting interactions with choice of dimension reduction approach, classification method, and data type. Dimension reduction approach, classification method, and data type seem to be the most influential factors on overall classification performance, and so it may be worthwhile to

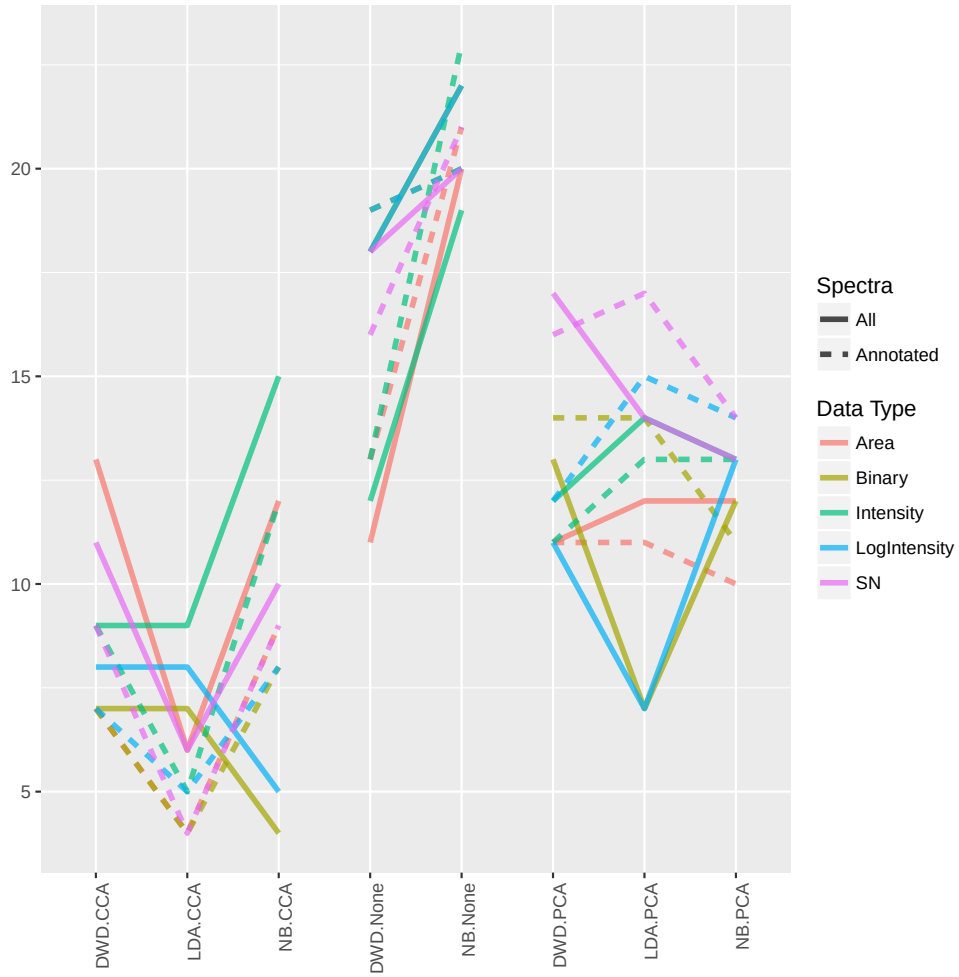


Figure 5.3: **Classification With/ Without Restricting to only Cancer Annotated Spectra.** LOO misclassification on the y-axis vs. the combination of classification and dimension reduction method used on the x-axis. The results from using all spectra and using only annotated tumour spectra are identified by use of solid and dashed lines respectively. The results from using each data type are identified by a single colour. In cases that include a dimension reduction step (PCA or CCA), results are only shown for the optimal choice for the number of dimensions, k, that is the k that achieves the lowest LOO misclassification. In cases when there are multiple k that achieve equal lowest LOO misclassification, we choose the smallest — the most parsimonous.

investigate these interactions further by designing new experiments with this goal specifically in mind. As it is, however, these data are inconclusive as to the effects restricting to cancer annotated spectra have on classification.

5.3.2 Binary Data

For the binary data, we consider the spatial smooth discussed in Section 2.5 prior to dimension reduction and classification. Figure 5.4 shows the results on the smoothed binary data, and of particular note Figure 5.4 shows that when we restrict to only cancer annotated spectra and apply either level of smoothing (0.15 or 0.25), CCA-LDA achieved the best LOO misclassification, one, that we have seen so far. Overall, smoothing seems to improve performance when combined with CCA-based variable selection, but does not show any such clear trend of improvement when used on the data with no dimension reduction or the PCA dimension reduced data.

5.3.3 Non-Binary Data

For the non-binary data types, we apply the normalisation as described in Section 4.4 in an attempt to reduce the unwanted variability in the non-binary measures of peak presence. The effect of normalisation on these results is shown in Figure 5.5. Figure 5.5 shows that normalisation does seem to have a clear overall effect on the LOO misclassification. However, one case of interest is when the normalised log-intensity data are used with CCA-LDA, including all spectra. This case achieves the equal best LOO misclassification of one observed so far. The other two results that achieved a LOO misclassification of one corresponded to the use of the binary data with smoothing, mentioned in the previous section.

The second preprocessing option specific to the non-binary data that we will consider relates to the averaging step — in which spectra from each patient are averaged on a per m/z bin basis. For non-binary data types we need to make a decision about how we treat missing peaks. So far, we have included zeroes for missing peaks when averaging spectra, but alternatively we could restrict the averaging to only spectra that have peaks. We now compare these two ways of taking averages/ treating missing peaks. Figure 5.5 shows 128 cases corresponding to every combination of non-binary data type, dimension reduction approach (CCA, no dimension reduction, or PCA), classification method (DWD, LDA, and NB), spectra restriction (all spectra or only annotated cancer spectra), and normalisation (with or without). We include zeroes for absent peaks when averaging in all 128 of the cases shown in Figure 5.5. Figure 5.6 shows these same 128 cases, but without including zeros for absent peaks when averaging. Figure 5.6 shows a similar lack of overall trend with respect to the effect of normalisation as in Figure 5.5.

Ignoring absent peaks when averaging (comparing Figure 5.5 and Figure 5.6) also does not seem to have an overall effect on classification performance, although it does seem to dramatically effect the classification in some specific cases. In particular, ignoring absent peaks when averaging the CCA variable selected log-intensity data produces remarkably low misclassifications. Specifically,

- in five of these cases a LOO misclassification of zero is achieved,
  - four of which correspond to the use of NB or DWD with or without restricting to annotated tumour tissue spectra, and without normalisation,
  - the fifth corresponds to the use of NB on the normalised data without restricting to annotated tissue.

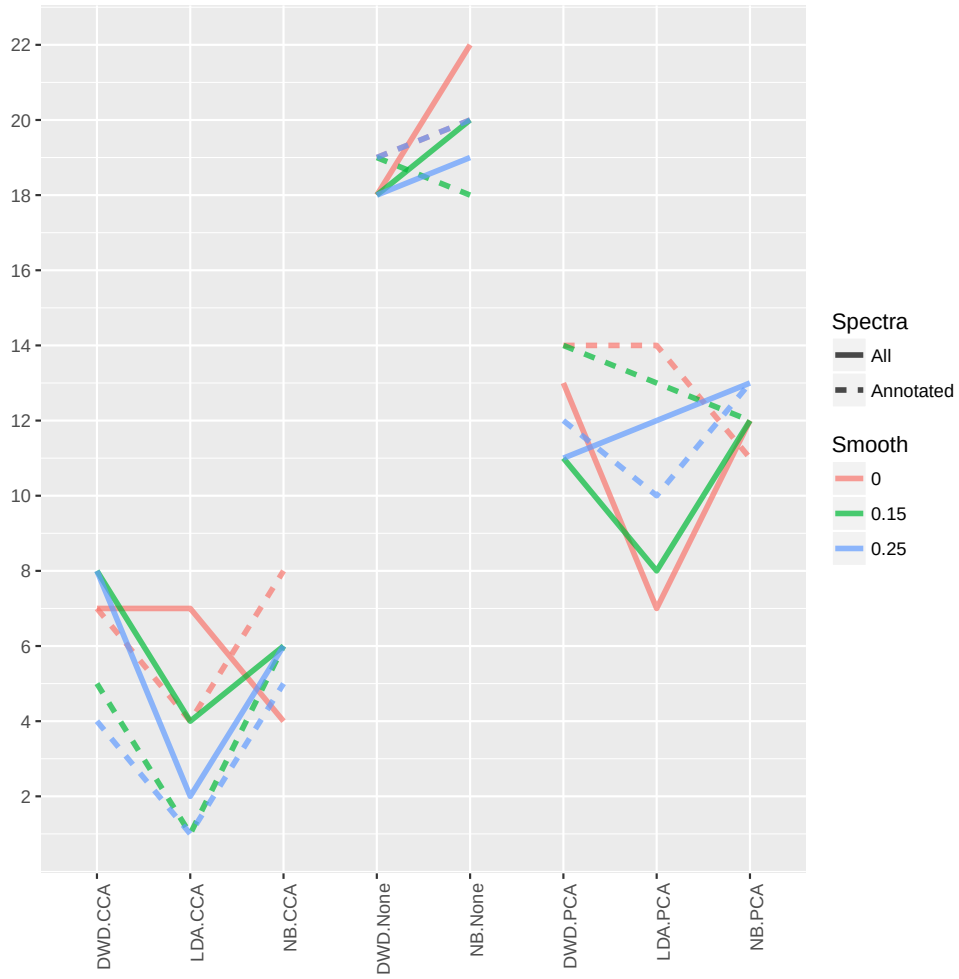


Figure 5.4: **Classification of Binary Data With/ Without Spatial Smoothing.** LOO misclassification on the y-axis vs. the combination of classification and dimension reduction method selected on the x-axis. The results from using all spectra and using only annotated tumour spectra are identified by use of solid and dashed lines respectively. The results from using no smoothing ($\tau = 0$), weak smoothing ($\tau = 0.15$), or medium smoothing ($\tau = 0.25$) are identified with colours. The smoothing is described in Section 2.5. In cases that include a dimension reduction step (PCA or CCA), results are only shown for the optimal choice for the number of dimensions, k, that is the k that achieves the lowest LOO misclassification. In cases when there are multiple k that achieve equal lowest LOO misclassification, we choose the smallest — the most parsimonous.

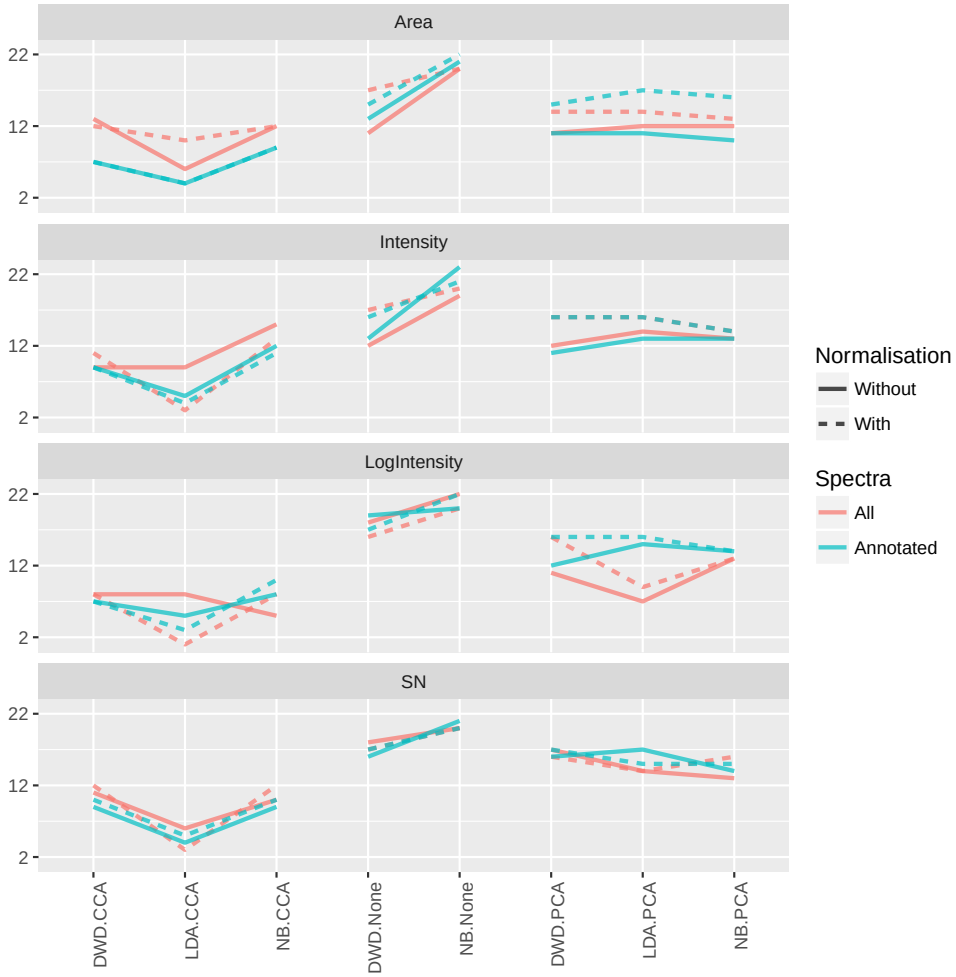


Figure 5.5: **Classification of Non-Binary Data With/ Without Normalisation — Part 1: Including Zeroes for Missing Values.** LOO misclassification on the y-axis vs. the combination of classification and dimension reduction method selected on the x-axis. The results from using each data type are shown in seperate panels. The results from using all spectra and using only annotated tumour spectra are identified by two colours respectively accross panels. The results from not using/ using normalisation are identified by use of solid and dashed lines respectively. All results shown include zeros for absent peaks when averaging. In cases that include a dimension reduction step (PCA or CCA), results are only shown for the optimal choice for the number of dimensions, k, that is the k that achieves the lowest LOO misclassification. In cases when there are multiple k that achieve equal lowest LOO misclassification, we choose the smallest — the most parsimonous.

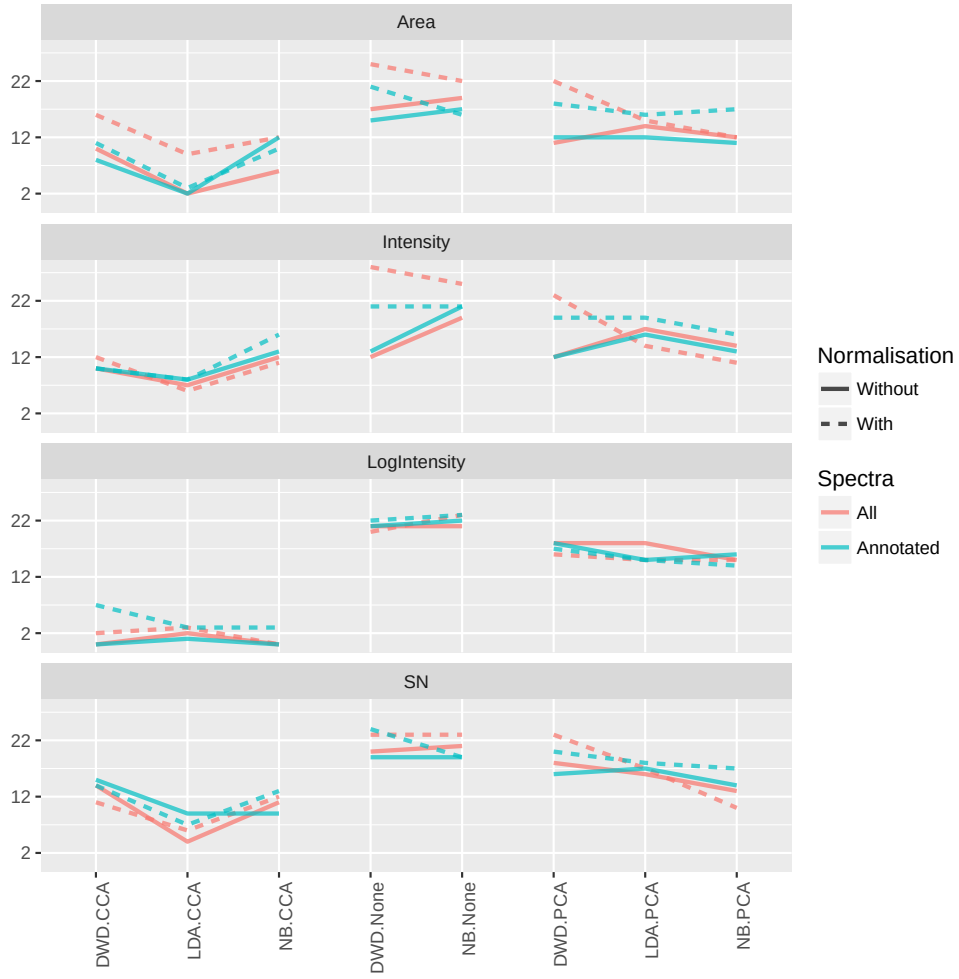


Figure 5.6: **Classification of Non-Binary Data With/ Without Normalisation — Part 2: Not Including Missing Values.** LOO misclassification on the y-axis vs. the combination of classification and dimension reduction method selected on the x-axis. The results from using each data type are shown in seperate panels. The results from using all spectra and using only annotated tumour spectra are identified by two colours respectively accross panels. The results from not using/ using normalisation are identified by use of solid and dashed lines respectively. All results shown do not include zeros for absent peaks when averaging. In cases that include a dimension reduction step (PCA or CCA), results are only shown for the optimal choice for the number of dimensions, k, that is the k that achieves the lowest LOO misclassification. In cases when there are multiple k that achieve equal lowest LOO misclassification, we choose the smallest — the most parsimonous.

- LDA achieves a LOO misclassification of one on these data without normalisation and restricting to only annotated spectra.

The fact that ignoring absent peaks when averaging produces so many extremely low LOO misclassification results warrants some further investigation. In Section 5.4 we consider the cases which achieved the best classification performance in more detail, including the six cases mentioned in the dot points above as well as the three other cases mentioned earlier. The preprocessing options used to achieve these nine cases are listed in Table 5.1.

5.4 The Lowest Misclassification Results

In Section 5.3 we presented a total of 304 classification results, represented across Figures 5.3-5.6, with some of these results being repeated from Figure 5.1 and Figure 5.2 respectively. These 304 cases result from every possible combination of the following options:

- **Dimension reduction approach** (PCA, CCA, or no dimension reduction),
- **Classification method** (NB, LDA, or DWD), *Note: LDA cannot be used if no dimension reduction is performed, as discussed in Section 4.1.*
- **Spectra included in patient-averages** (all, or only annotated tumour spectra),
- **Data type** (area, binary, intensity, log-intensity, or SNR),
  - When non-binary data types are used,
    - **Normalisation** (with, or without) as described in Section 4.4, and
    - **Treatment of absent peaks when averaging** (include as zeros, or ignore),
  - When binary data is used, **Spatial smoothing** ($\tau = 0$, $\tau = 0.15$, or $\tau = 0.25$) as described in Section 2.5,

Table 5.1: Classification results that achieve a LOO misclassification of zero or one. All results shown achieved said LOO misclassification using CCA-based variable selection, but we include the corresponding results when PCA and no dimension reduction are used for comparison. In the cases when dimension reduction is used, the number of dimensions reduced too, k, is also included. The 'ID' column is used to identify these results with those shown in Figure 5.7. Abbreviations follow. Norm: Normalisation. Annot: Annotated. log(I): log-intensity. W/o: Without.

Method	Spectra	Data Type	τ	Norm	Absent Peaks	LOO	PCA LOO	PCA k	CCA LOO	CCA k	ID
DWD	All	log(I)		W/o	Ignore	21	18	5	0	23	1
DWD	Annot	log(I)		W/o	Ignore	21	18	35	0	14	2
LDA	Annot	Binary	0.15				13	36	1	31	3
LDA	Annot	Binary	0.25				10	37	1	23	4
LDA	Annot	log(I)		W/o	Ignore		15	33	1	26	5
LDA	All	log(I)		With	Include		9	39	1	24	6
NB	All	log(I)		W/o	Ignore	21	15	30	0	42	7
NB	Annot	log(I)		W/o	Ignore	22	16	20	0	19	8
NB	All	log(I)		With	Ignore	23	15	37	0	31	9

Of these 304 results, 9 achieve a LOO misclassification of zero or one, and these are shown in Table 5.1. Each of these results was noted in the discussion of Section 5.3. The only option that is the same across all 9 of these best results is that CCA-based variable selection was used. The results corresponding to using all the same options except with alternative dimension reduction approaches are also included in Table 5.1 for comparison. Such a comparison reveals the dramatic improvement achieved by CCA-based variable selection compared to either PCA dimension reduction or no dimension reduction in these cases. None of the 9 cases shown in Table 5.1 achieve a LOO misclassification below 20 when no dimension reduction is done prior to classification — out of a total of 43 patients, this is not much better than coin tossing. PCA dimension reduction performs slightly better than no dimension reduction — achieving LOO misclassifications as low as 9 and 10. As previously mentioned, all these cases achieve LOO misclassifications of zero or one using CCA-based variable selection. As discussed in Section 5.2, it is perhaps not entirely surprising that CCA-based variable selection outperforms these other approaches, as it is the only 'supervised' approach to dimension reduction we have considered — i.e. it takes into account information about the class labels — but the degree to which it outperforms these other approaches is still surprising. Surprise aside, this supports the conclusion we made in Winderbaum et al. (2016) — that one of the most important factors in determining classification performance is the approach taken to dimension reduction.

In cases where dimension reduction is used the number of dimensions to reduce too, k, needs to be chosen and, as discussed in Section 5.2, we consider only the 'optimal' k — the smallest k which minimises the LOO misclassification, included in Table 5.1. This simplification allows for us to represent the results from more permutations and variations in a single plot, and has been useful in Section 5.3. However now that we are interested in a relatively smaller number of cases we can consider these cases in more detail by varying k. Figure 5.7 shows the LOO misclassification as the number of CCA ranked variables selected k is varied, similarly to Figure 5.2, for the 9 best sets of options as listed in Table 5.1. Figure 5.7 shows the interesting pattern that the DWD and NB LOO misclassification tend to reduce and stabilise at a minimum value as k increases, but the LDA LOO misclassification seems to achieve a local minima somewhere in the range $20 < k < 30$, rising again for $k > 30$ — note that there is one possible exception to this trend (ID = 3). In a traditional low-dimensional setting, it is typical to expect adding variables to improve classification. However in HDLSS data such as this, it is not unusual to observe that once a certain number of dimensions is reached, here around the $20 - 30$ range, additional variables can begin to behave as noise and worsen results. This change in behaviour allows for 'optimal' choices to be made for the number of dimensions to reduce too.

Another advantage of the CCA variable selection approach is that it selects from existing variables, thus preserving their interpretation as m/z bins. Thus, we can look into the rankings that produced these best results, and find which m/z values are highly ranked, as these could be potential targets for follow up validation studies. As choice of classification method does not influence the variable ranking, there are 6 unique sets of preprocessing options represented amongst the 9 results of Table 5.1. Each of these 6 sets of preprocessing options will produce different rankings, and each of these will have three rankings produced from the three shifted-bin analyses as discussed in Section 4.2. Parallel LC-MS analyses were conducted on tissue from the endometrial TMAs for protein identifications, so that these identifications could then be matched to these highly ranked m/z bins in order to infer proteins

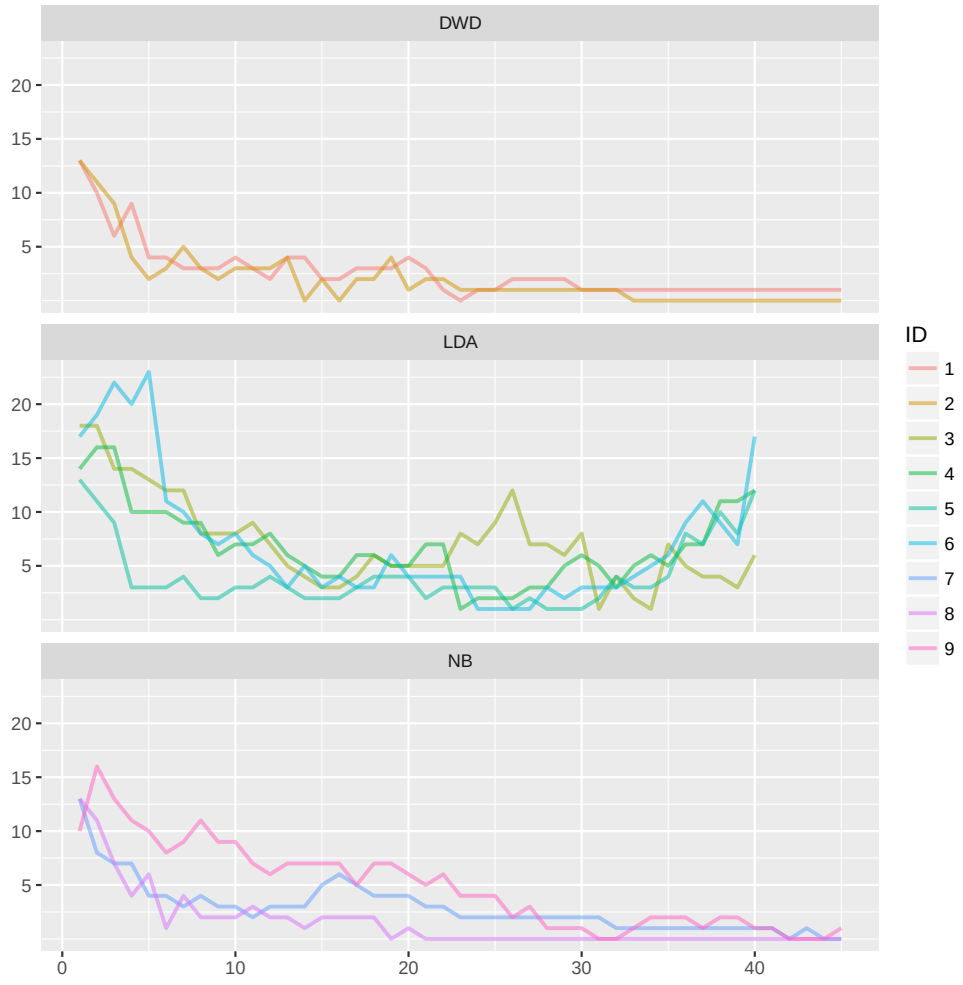


Figure 5.7: LOO misclassification on the y-axis vs the number of variables, k, on the x-axis where k is the number of variables retained in the CCA-based variable selection step. Results using each of the combinations of preprocessing options listed in Table 5.1 are shown. Each of these combinations of preprocessing options are identified by ID and colour, and separated into panels by classification method.

that could be important to the classification of LNM status. Several proteins of interest were identified through this matching, and follow up validation studies are currently being undertaken to further investigate the link between these proteins and LNM in endometrial cancer. As an example, one of the most consistently recurring m/z values is that centred around $m/z = 1198.701$, which is ranked in the top 20 variables for 6 of the 18 rankings. The most likely parent protein identified from the LC-MS for this peptide m/z is an Actin, most likely aortic smooth muscle Actin (UniProtKB entry name: `ACTA_HUMAN`). Two other masses likely to be Actin peptides are also highly ranked in several of the rankings, specifically those centred around $m/z = 1161.565$ and $m/z = 1501.749$. Table 5.2 shows all potential matches from the LC-MS identifications to these three m/z values, notice that Actin is not the only possible parent protein — there are other possibilities, and these are also being pursued in follow-up work.

Table 5.2: LC matching to Actin peptides

MALDI m/z	UniProtKB Entry Name	Peptide Sequence	MASCOT expect	error (m/z)	error (ppm)
1198.70	ACTC_HUMAN	AVFPSIVGRPR	0.00	-0.01	-4.25
	H14_HUMAN	ASGPPVSELITK	0.00	0.03	27.21
	ACTA_HUMAN	AVFPSIVGRPR	0.00	-0.01	-4.25
	ACTB_HUMAN	AVFPSIVGRPR	0.00	-0.01	-4.25
	H12_HUMAN	ASGPPVSELITK	0.00	0.03	27.21
	ACTS_HUMAN	AVFPSIVGRPR	0.00	-0.01	-4.25
	RO60_HUMAN	LGLENAEALIR	0.00	0.02	17.86
1161.56	CUL5_HUMAN	TLWSLVAFPK	0.00	-0.10	-88.15
	ACTS_HUMAN	EITALAPSTMK	0.00	-0.05	-46.57
	RET4_HUMAN	FSGTWYAMAK	0.01	0.02	21.09
	ACTA_HUMAN	EITALAPSTMK	0.00	-0.05	-46.57
	ACTB_HUMAN	EITALAPSTMK	0.00	-0.05	-46.57
	ACTC_HUMAN	EITALAPSTMK	0.00	-0.05	-46.57
1501.75	FIBA_HUMAN	MELERPGGNEITR	0.00	0.01	4.17
	SPTN1_HUMAN	EANELQQWINEK	0.00	0.02	13.90
	IF4A1_HUMAN	GFKDQIYDIFQK	0.00	-0.02	-13.07
	ACTA_HUMAN	IWHHSFYNELR	0.00	0.02	10.30
	MYO1C_HUMAN	MSLLQLVEILQSK	0.02	-0.12	-77.79
	FA49B_HUMAN	MSLFYAEATPMLK	0.00	0.01	4.04

5.5 Measuring Stability/ Overfitting/ Leverage

So far in this chapter, and in particular in Section 5.3, we have considered the results of classification using a considerable array of different options. We discuss the cases that resulted in the best LOO misclassification in Section 5.4. Due to the nature of our approach — applying many different combinations of processing options — it can be difficult to judge if these best results are due to accurate prediction, or random chance combined with a sufficiently large number of permutations of processing options.

I introduce a heuristic measure of classification rule stability in Equation 5.1, comparable to the ideas of leverage from regression (Everitt and Skrondal, 2002). Leverage measures the effect that removal of a single observation has on the parameter estimates in regression. Analogously the heuristic we introduce below is a measure of the effect that removing a single observation has on the direction vector $\boldsymbol{d}$

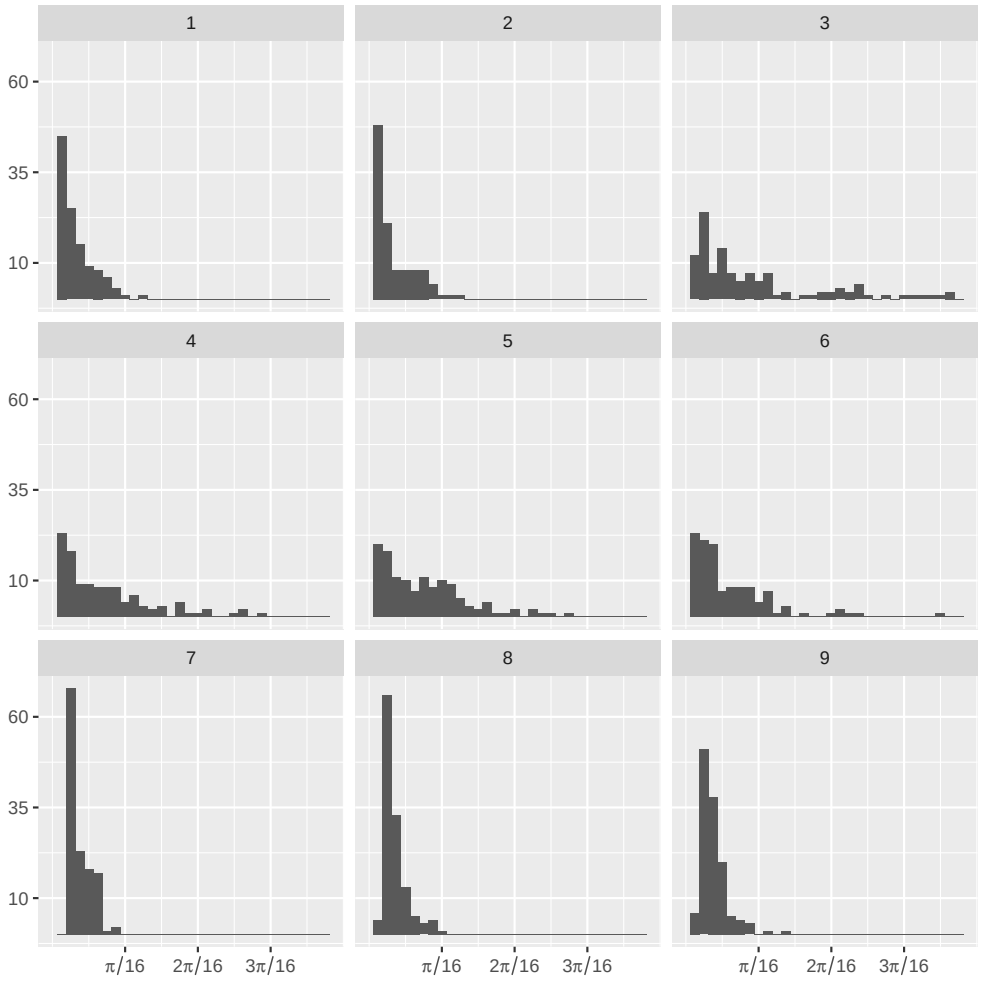


Figure 5.8: Frequency histograms of stability heuristic values from the 9 CCA variable selection cases shown in Table 5.1. Each histogram is presented in a seperate panel labeled 1-9 corresponding to the ID column of Table 5.1. Each histogram consists of 129 stability heuristic values — one for each of the 43 patients for each of the three shifted-bin analyses.

as in the formulation of Equation 4.1 for a linear classification rule trained from the data. This idea also naturally follows from the ideas of LOO CV, where the analysis is repeated with each observation removed. If we let $\boldsymbol{d}$ be the direction vector of the classification rule trained from all the data, and let $\boldsymbol{d}_i$ be the direction vector of the classification rule trained from the data with the ith observation removed, then the heuristic we will consider is the internal angle between the two vectors,

$$\arccos\left(\frac{\boldsymbol{d}\cdot\boldsymbol{d}_i}{|\boldsymbol{d}||\boldsymbol{d}_i|}\right) = \arccos\left(\boldsymbol{d}\cdot\boldsymbol{d}_i\right) \quad \text{as typically} \quad |\boldsymbol{d}| = |\boldsymbol{d}_i| = 1, \tag{5.1}$$

which is directly related to the cosine distance of Definition 2. The value of the heuristic defined in Equation 5.1 can be interpreted as an angle, indicating the change in direction of the trained classification rule when the ith observation is removed. An angle of zero indicates there is no difference between the original direction vector and the direction vector trained on the data with the ith observation removed. Larger values of this heuristic indicate larger changes in the direction vector — larger angles between the two direction vectors.

Although the heuristic of Equation 5.1 does not entirely address the difficulty mentioned above in judging if results are due to accurate prediction or random chance, it can nonetheless provide some insight into the stability or sensitivity of these classification methods to small changes in the data — specifically, to the removal of individual observations. Figure 5.8 shows a histogram of these heuristic values calculated for the 9 cases of Table 5.1. Each of the 9 results of Table 5.1 actually consist of three shifted-bin analyses, and so each of the 9 histograms represents 129 heuristic values, one for each of the 43 patients for each of the three shifted-bin analyses.

It can be seen from Figure 5.8 that all these cases have strongly right-skewed distributions with the majority of values falling very close to zero. This is good as it indicates that overall in the majority of these cases, the direction vector is not very sensitive to the removal of individual observations from the training dataset. Some notes on Figure 5.8:

- Three distinct distributions can be observed corresponding to the three classification methods: DWD (ID = 1-2), LDA (ID = 3-6), and NB (ID = 7-9).

- The distributions for the LDA cases have significantly thicker tails than the other cases — seeming to demonstrate the most overall instability of the three methods. LDA is also the only method of the three to show heuristic values above $\frac{\pi}{8}$ radians (22.5 degrees), although only for a very small number of values.

- The distributions of the DWD and NB cases show similar thickness tails, but different modes. The DWD cases tend to have modes below $\frac{\pi}{64}$ radians, while the NB cases have modes above $\frac{\pi}{64}$ radians and infact do not have any heuristic values below $\frac{\pi}{128}$ radians at all.

It is interesting to speculate on the possible link between these stability heuristic results, and the results shown in Figure 5.2, in which it was noted that DWD showed less variance in LOO misclassification as the number of dimensions k was varied. Furthermore, there may be reason so consider classification methods other than LDA — despite the fact that LDA tends to achieve the best LOO misclassification results — as LDA is also the most sensitive to small changes in the data according to this heuristic, which is an undesirable property in a classification rule.

5.6 Conclusions

As summarised at the beginning of Section 5.4, we have presented results representing a considerable array of different options for classification in Section 5.3. For convenience we repeat the options which we have considered:

- **Dimension reduction approach** (PCA, CCA, or no dimension reduction),
- **Classification method** (NB, LDA, or DWD), *Note: LDA cannot be used if no dimension reduction is performed, as discussed in Section 4.1.*
- **Spectra included in patient-averages** (all, or only annotated tumour spectra),
- **Data type** (area, binary, intensity, log-intensity, or SNR),
  - When non-binary data types are used,
    - **Normalisation** (with, or without) as described in Section 4.4, and
    - **Treatment of absent peaks when averaging** (include as zeros, or ignore),
  - When binary data are used, **Spatial smoothing** ($\tau = 0$, $\tau = 0.15$, or $\tau = 0.25$) as described in Section 2.5,

Amongst these results, there are some general trends and suggestions that can be made based on these trends. Ideally we would like to be able to make recommendations on approaches, methods, and preprocessing options that tend to be more effective when classifying MALDI-MSI TMA data. However, of the results discussed in this chapter, no single set of decisions seems to demonstrate superior results over all other options. The best strategy may be to try several options and use whichever performs best in any given circumstance. That said, some decisions had a more pronounced effect on classification performance than others. Summarising and discussing these effects is the focus of this section. Overall, the factor that seems to have the biggest effect on classification performance is the **Dimension reduction approach** taken, with CCA variable ranking performing very well. **Classification method** and **Data type** also seemed to have significant effects on classification performance, and in particular seemed to have strong interaction effects with each other and choice of **Dimension reduction approach**. The remaining preprocessing variants we considered did not seem to have a consistent effect on classification performance. These options are: **Spectra included in patient-averages** (all or only annotated tumour spectra), **Spatial Smoothing** of the binary data, **Normalisation** and **Treatment of absent peaks when averaging** in the non-binary data.

Furthermore, it is also of interest if these trends extend to classification of MALDI-MSI TMA data in general, or if they are artefacts of the endometrial data we have considered here. In order to investigate this possibility we have replicated all the analyses in this chapter using a different dataset — relating to vulvar cancer, as mentioned briefly at the end of Section 1.5.3. The results of classification on the vulvar data are included as Appendix D, in which Figures D.1-D.6 mirror Figures 5.1-5.6 in a one-to-one fashion but for the vulvar data.

Dimension Reduction Approach

First and foremost, the clearest trend and strongest conclusion from these results is that the CCA-based variable ranking performs very well. The importance of dimension reduction, and the superiority of the CCA-based variable ranking approach are also very clear in the vulvar data analyses of Appendix D. Furthermore, the CCA-based variable ranking method has the additional (very significant) advantage of interpretability — selected variables correspond directly to analytes of biological interest, as discussed in Section 5.4.

Classification Method

Overall, CCA-LDA seemed to achieve the best results in most circumstances, with a few exceptions noted in Table 5.1. However based on our heuristic stability analysis of Section 5.5, LDA showed the worst stability. This instability of the LDA method could contribute to higher variance in LOO misclassification, and this could potentially explain the better results achieved with LOO. The results discussed are largely those chosen from a range of possible dimension-reductions, with the dimension-reduction chosen such that the minimum LOO misclassification is achieved. If the LDA method is more sensitive to small changes in the data, as the results of the heuristic stability analysis in Section 5.5 suggest, this could mean the results would vary more as the number of dimensions used is varied. Higher variance in the LOO misclassification combined with the 'optimality selection' of choosing the dimension that minimises the LOO misclassification could be biasing the results to show LDA performing better. This raises questions of how to appropriately measure the performance of classification methods in such cases, but does not conclusively answer any such questions. DWD demonstrated the best stability in the heuristic analysis of Section 5.5, and this may indicate that more consistent results could be achieved with DWD. It is difficult to say with any certainty either way, and so although choice of classification method is clearly very important, it is nonetheless difficult to recommend a single classification method as being superior to all others. Instead we recommend using several options and selecting that which performs best or combining the results of a number of good options in a sensible way.

Data Type

Overall, the log-intensity data achieved very good LOO misclassification results, with the binary data also achieving some notable local optimums, as noted in Table 5.1. This is also true of the vulvar data results in Appendix D. As such we suggest that the log-intensity data serves as a good starting point for classification of these data, but exploration of alternative data types, including the binary data, may also yield improvement and should be pursued if optimising results is of interest.

The distribution of intensity values in a typical MALDI-glsims dataset well approximates an exponential distribution, which could contribute to why the log-intensity values achieve good results using these linear classification methods. It should be noted that other non-binary data types, such as SNR, also follow a similar exponential decay distribution of values typically, and so considering their log-transformed analogues may also be of interest.

Cancer Annotation

In principle, restricting to a single tissue type should reduce the within-patient variability and thereby facilitate more accurate prediction. However, the results do not support this hypothesis — showing no consistent effect to this restriction. This same lack of consistent trend is apparent in the vulvar cancer results of Appendix D. There are several possible, not mutually exclusive, explanations that could account for this, including:

- Restricting to annotated spectra reduces the total amount of spectra used in the analyses, and this could lead to more noisy patient-averages as each average is obtained from a smaller number of observations (spectra). It is possible that restricting to annotated spectra does reduce the variability in the data by restricting to a single tissue type, but that reducing the total number of spectra also increases the variability of the averages, and these two competing effects cancel each other out, resulting in no net effect on the classification performance.

- There may exist characteristics of the surrounding non-tumour stroma tissue that are important in the prediction of LNM, and that this information is lost when restricting to only tumour tissue. Similarly this could compete with the effect of reducing the variability due to multiple tissue types being considered, and result in no net change being observable. There is some evidence to support this hypothesis, specifically Oppenheimer et al. (2010) demonstrated that tissue adjacent to a tumour, histologically classified as non-tumour, can share molecular characteristics with the tumour tissue. Oppenheimer et al. (2010) suggested that this phenomena could be involved in tumour recurrence post resection, but the same phenomena could also be involved in explaining why restricting to histologically annotated tumour regions does not produce a consistent improvement in classification performance.

These points, particularly the second dot point above, warrant further investigation in future research.

Smoothing

In most of the cases considered for the binary data, and particularly those cases that achieved the best results, it seems that spatial smoothing has a good net effect on LOO misclassification. This is unsurprising, as the smoothing should reduce the noise, and thus allow signals to be more easily detected. For future research, pursuing spatial smoothing techniques for the non-binary data, such as simple kernel density smoothing, could be of interest.

Normalisation

It makes intuitive sense that normalisation should reduce the variability in the data, facilitating more accurate classification. However, the results do not support this — LOO misclassification shows no obvious trend related to the use of normalisation prior to classification. Infact, overall, classification performance tends to be worse when normalisation is used. This worsening of the classification performance when normalisation is used can, for example, be seen in the results using area and intensity data in Figures 5.5 and 5.6. Despite this overall trend however, some of the results achieving the overall best LOO misclassification shown in Table 5.1 include

use of normalisation. One possible explanation for this seeming contradiction is that using normalisation could introduce additional degrees of freedom in preprocessing decisions, and this could allow for the 'optimality selection' bias effect to find better minima. We discussed this optimality bias effect previously, in relation to classification methods and the heuristic stability measure. Further investigation could be of use in elucidating explanations for this behaviour, but ultimately our results are inconclusive on the effect of normalisation.

Absent Peaks

Similarly to normalisation, the overall trend seems to be that ignoring absent peaks when averaging worsens LOO misclassification more often that not, but the minority of results contradicting this trend achieve some of the overall best LOO misclassification results (shown in Table 5.1). However, this trend is not replicated in the vulvar cancer results shown in Appendix D. In the vulvar cancer results of Appendix D, ignoring absent peaks has no obvious net effect on classification performance, suggesting that the minor downward trend in the endometrial cancer results may simply be an artefact of these data. Another possible explanation for this downward trend is that considering multiple options for how to treat absent peaks when averaging introduces an additional degree of freedom in the preprocessing decisions considered, and thereby contributes to the 'optimality bias' as discussed above in regards to normalisation. Similarly to the normalisation, these hypotheses warrant further investigation, but ultimately our results are inconclusive on the effect of ignoring absent peaks when averaging.

Concluding Remarks

MALDI-MSI has two strengths that we have focussed on: preserving spatial information, and facilitating the classification of clinically relevant diagnostic and prognostic factors through the use of TMAs.

In Chapters 2 and 3 we explored the spatial aspect of MALDI-MSI data. In Chapter 2 we demonstrated that cancerous tumour tissue could be seperated from its surrounding non-tumour tissues by using an automated clustering approach and that this seperation could be used to implement a DIPPS-feature selection scheme for selecting a short-list of peptides that are more highly expressed in tumour tissue than non-tumour tissue. In Section 3.1 we explored the use of these short-lists in gaining information about the within-patient and between-patient variability in MALDI-MSI data, conluding that although this variability was significant, considering technical replicates can allow for biologically significant differences between patients to be detected reproducibly. In Section 3.2 we demonstrated that our DIPPS-feature selection approach can be applied in another way — to find glycan signals (Gustafsson et al., 2015). We published this DIPPS-feature selection approach as Winderbaum et al. (2015).

In Chapters 4 and 5 we consider the classification of MALDI-MSI TMA data. In Chapter 4 we introduced and disucssed methods for the classification of such MALDI-MSI TMA data. In Chapter 5 we applied these methods to classify LNM in the endometrial data of Mittal et al. (2016), and considered various options for pre-processing, dimension-reduction, and classification of these data — comparing the classification performance for each option. Ultimately we concluded that the most important factor to classification performance overall was the approach taken to dimension reduction — with CCA-based variable selection performing very well. Some options also seemed to have very strong interaction effects with each other — resulting in particular combinations of choices having much improved classification performance despite each of the individual options not having big effects in general. These interaction effects ultimately led to the overall conclusion that the most consistent set of options was using CCA-LDA on the log-intensity data. In addition to these results from the endometrial data we also replicated our analysis on a second dataset relating to vulvar cancer, as shown in Appendix D. Many of the more minor results from the endometrial data were not reproduced in the vulvar data. However the most obvious trends, including the consistently good classification performance of CCA-LDA on the log-intensity data, were the same in the vulvar data, strengthening these results. We published our CCA-based variable selection approach with these results as Winderbaum et al. (2016).

In addition to developing a framework for feature selection and classification of MALDI-MSI data, this thesis contains new mathematical and statistical results, namely:

- The spatial smooth of Section 2.5, which applies not only to proteomics data analysis, but to binary data in general.

- The analytic form for the matrix inverse derived in Appendix C, which is of interest not only in linear regression, but is infact a result for a more general family of matrices.

Appendices

Appendix A

Binning

Here we introduce details on the binning algorithm used, including notation and definitions. After introducing binning (Section A.1) and some directly related concepts (Section A.2 and Section A.3), I go on to discuss the binary / summed binary data equivalence (Section A.4) which is related to choice of bin size, and important when considering non-binary (such as intensity) data. Considering binning using alternative bin locations will also be relevant, as choice of bin location is arbitrary, and the discussion of (Section A.5) provides a framework within which sensitivity of results to bin location can be explored.

A.1 Binning Algorithm for Peaklist Data

The binning method considered here could be used on any functional data where 'features' have been identified. To avoid ambiguity I define binning here explicitly in the context of peaklist data, but all the concepts involved are completely general. As mentioned in Section 1.5, data we consider will be in the pre-processed 'peaklist' format — meaning the data can be represented as a list of peaks, each with an associated m/z value and parent spectrum as well as other properties. If we denote the m/z value of the ith peak associated to the jth spectrum m_{ij}, then we introduce notation in Equation A.1 for the maximum and minimum m/z values in a dataset, i.e.

$$m_{\min} = \min_{i,j} \{m_{ij}\} \quad \text{and} \quad m_{\max} = \max_{i,j} \{m_{ij}\}. \tag{A.1}$$

In Equation A.2 we introduce notation for n_{first} - the number of adjacent (non-overlapping) bins of size b that lie between 0 and $m_{\min}$, and n_{last} - the number of adjacent (non-overlapping) bins of size b needed to cover both 0 and $m_{\max}$, specifically

$$n_{\text{first}} = \left\lfloor \frac{m_{\min}}{b} \right\rfloor \quad \text{and} \quad n_{\text{last}} = \left\lceil \frac{m_{\max}}{b} \right\rceil. \tag{A.2}$$

The notation introduced in Equation A.1 and Equation A.2 is sufficient to define Algorithm A.1, which explicitly defines the process of producing binned data and is illustrated in Figure A.1.

Algorithm A.1. Binning: *Given a bin size $b > 0$ and a dataset consisting of n spectra in peaklist format, using the notation introduced in Equation A.1 and Equation A.2,*

1. *Construct $n_{last} - n_{first} + 1$ intervals (bins) with left endpoints open, right closed of width b and with centres*

$$n_{first}b, \;\; (n_{first}+1)b, \;\; (n_{first}+2)b, \;\; \ldots, \;\; n_{last}b.$$

2. *Use the bins from the previous step to produce a* $(n_{last} - n_{first} + 1) \times 1$ *vector* $\boldsymbol{x}_{\bullet j}$ *for each spectrum* $j = 1, 2, \ldots, n$ *where the* $\boldsymbol{x}_{\bullet j}$ *are one of either*

   - ***Binary Data:*** *The* $\boldsymbol{x}_{\bullet j}$ *are such that for each* j*, the* i^{th} *entry of* $\boldsymbol{x}_{\bullet j}$ *is zero if spectrum* j *has no peaks in the the* i^{th} *bin, or one if spectrum* j *has at least one peak in the the* i^{th} *bin.*
   - ***Summed Binary Data****: The* $\boldsymbol{x}_{\bullet j}$ *are such that for each* j*, the* i^{th} *entry of* $\boldsymbol{x}_{\bullet j}$ *is* k *if spectrum* j *has exactly* k *peaks in the the* i^{th} *bin.*

3. *Construct a* $d \times n$ *data matrix* $\mathbb{X}$ *(where* $d = n_{last} - n_{first} + 1$*) whose columns are the* $\boldsymbol{x}_{\bullet j}$*.*

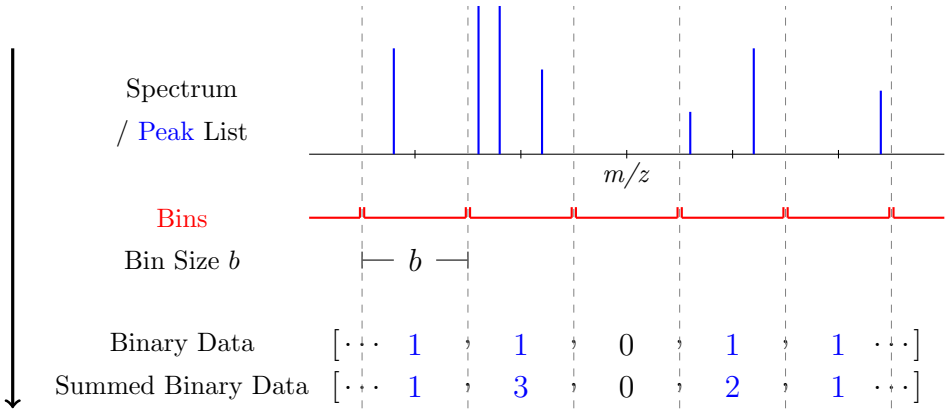


Figure A.1: Schematic illustrating the binning process (Algorithm A.1). Bins are used to partition the continuous m/z range, and peaks are identified by the bin within which they occur. Moving from the top of the figure down, peaklist data can then be converted into either binary, or summed binary data by constructing a vector whose entries are respectively either; indicators for, or counts of, the number of peaks in the corresponding bin. In Algorithm A.1 a vector is constructed for each spectra, and these are concatenated into a data matrix as columns.

When I write "a $d \times n$ binary (binned) data matrix $\mathbb{X}$" or "a $d \times n$ summed binary (binned) data matrix $\mathbb{X}$" I refer to a data matrix $\mathbb{X}$ of either binary, or summed binary data respectively - as produced by Algorithm A.1 above. These binned data are used in the analysis of both the ovarian and endometrial cancer datasets described in Section 1.5.1 and Section 1.5.3 respectively.

A.2 Invariance Under Removal of Empty Bins

I will refer to bins that contain no peaks in any spectra as empty bins. Algorithm A.1 can (particularly for small bin sizes) produce empty bins. In a data matrix $\mathbb{X}$ produced by Algorithm A.1, each column corresponds to a spectrum, and each row corresponds to a bin. For reasons of computational speed it is often desirable to remove the rows of a data matrix $\mathbb{X}$ corresponding to empty bins, but we need to know what effect this will have on analyses.

In many cases removing empty bins has no effect on the results of further analyses. In this section I briefly discuss under which conditions removing empty bins will have no effect on further analyses, and some common examples of distances that are invariant under removal of empty bins.

Consider a $d \times n$ data matrix $\mathbb{X}$ whose jth column is denoted $\boldsymbol{x}_{\bullet j}$. Let d_{empty} be the number of empty rows of $\mathbb{X}$ — rows corresponding to empty bins across all spectra. We construct a new data matrix, $\mathbb{X}^*$ by removing the empty rows of $\mathbb{X}$ and let $\boldsymbol{x}^*_{\bullet j}$ denote the j^{th} column of $\mathbb{X}^*$.

Definition 17. Invariance under the removal of empty variables: *We call a distance D invariant to the removal of empty variables if*

$$D(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet k}) = D(\boldsymbol{x}^*_{\bullet j}, \boldsymbol{x}^*_{\bullet k}) \quad \forall\, j,k = 1,2,\ldots,n \quad \forall\, n,d \quad \textit{and} \quad \forall\, 0 \leq d_{empty} \leq d$$

Definition 17 holds for some pseudometrics, and not others. The Euclidean, cosine, and Hamming distances are examples of pseudometrics that are invariant under the removal of empty variables. In Definition 3 we define the Hamming distance as

$$D_{Ham} : \{0,1\}^d \times \{0,1\}^d \to [0,d], \qquad D_{Ham}(\boldsymbol{x},\boldsymbol{y}) = d - \boldsymbol{x}\cdot\boldsymbol{y} - (1-\boldsymbol{x})\cdot(1-\boldsymbol{y})$$

which can be interpreted as the number of positions in which the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ differ. An alternate definition for the Hamming distance is

$$D^*_{Ham} : \{0,1\}^d \times \{0,1\}^d \to [0,1], \qquad D^*_{Ham}(\boldsymbol{x},\boldsymbol{y}) = \frac{d - \boldsymbol{x}\cdot\boldsymbol{y} - (1-\boldsymbol{x})\cdot(1-\boldsymbol{y})}{d},$$

which can be interpreted as the proportion of positions that differ between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. This alternate definition for the Hamming distance, D^*_{Ham}, is the definition used in the MATLAB function `kmeans` via the `pdist` function, and is an example of a pseudometric that is not invariant under the removal of empty variables. When we use the term 'Hamming distance', we refer to Definition 3.

A.3 Matching Bins Between Datasets

I mentioned in the introduction to Section 2.1, an advantage of binning over data-driven methods such as those described in Section 3.2, is that comparisons of spectra within a single dataset can be extended to comparisons between multiple datasets in a natural and computationally efficient way. In this section I will explicitly define this natural extension of binning to comparisons between multiple datasets in Algorithm A.2, and briefly discuss the significance of invariance under the removal of empty variables (Definition 17) to these comparisons.

Let $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$ be $d_1 \times n_1$ and $d_2 \times n_2$ binned data matrices produced by Algorithm A.1 with some bin size b from two different peaklist datasets, which I will refer to as dataset (1) and dataset (2) respectively. If we wish to compare the two datasets, we would like their rows to correspond to the same m/z bins, which would allow a natural comparison of spectra from one dataset to spectra in the other. Algorithm A.2 describes how to modify these data matrices so that their rows correspond to the same m/z bins. I extend the notation introduced earlier by adding a superscript to denote dataset; Let $m^{(\nu)}_{ij}$ be the m/z location of the i^{th} peak in the jth spectrum of dataset (ν) for $\nu = 1,2$. Similarly, I extend the notation of Equation A.1 to get

$$m^{(\nu)}_{\min} = \min_{i,j}\left\{m^{(\nu)}_{ij}\right\} \quad \text{and} \quad m^{(\nu)}_{\max} = \max_{i,j}\left\{m^{(\nu)}_{ij}\right\} \quad \text{for } \nu = 1,2, \tag{A.3}$$

and the notation of Equation A.2 to get

$$n_{\text{first}}^{(\nu)} = \left\lfloor \frac{m_{\min}^{(\nu)}}{b} \right\rfloor \quad \text{and} \quad n_{\text{last}}^{(\nu)} = \left\lceil \frac{m_{\max}^{(\nu)}}{b} \right\rceil \quad \text{for } \nu = 1, 2 \tag{A.4}$$

similarly. Algorithm A.2 modifies $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$ by adding empty rows such that the rows of the modified matrices correspond to the same bins.

Algorithm A.2. Matching bins between two datasets: *Using the notation introduced in Equation A.3 and Equation A.4, and without loss of generality letting* $m_{min}^{(1)} \leq m_{min}^{(2)}$ *and* $m_{max}^{(1)} \leq m_{max}^{(2)}$*,*

1. *Modify* $\mathbb{X}^{(2)}$ *by adding* $n_{first}^{(2)} - n_{first}^{(1)}$ *empty rows to produce* $\begin{bmatrix} \mathbf{0}_{(n_{first}^{(2)} - n_{first}^{(1)}) \times d_1} \\ \mathbb{X}^{(2)} \end{bmatrix}$.

2. *Modify* $\mathbb{X}^{(1)}$ *by adding* $n_{last}^{(2)} - n_{last}^{(1)}$ *empty rows to produce* $\begin{bmatrix} \mathbb{X}^{(1)} \\ \mathbf{0}_{(n_{last}^{(2)} - n_{last}^{(1)}) \times d_1} \end{bmatrix}$.

If $\left(n_{first}^{(2)} - n_{first}^{(1)}\right)$ *or* $\left(n_{last}^{(2)} - n_{last}^{(1)}\right)$ *are zero, do not modify the data matrix in (*1.*) or (*2.*) respectively.*

The modified data matrices produced by Algorithm A.2 are comparable, as their rows correspond to the same m/z intervals. Comparisons between an arbitrary number of datasets is possible either by iterative use of Algorithm A.2 or a simple modification of Algorithm A.2 that involves the maximum and minimum m/z values across all the datasets considered.

Note that invariance under removal of constant/empty variables (Definition 17) is equivalent to invariance under the addition of finitely many empty variables. What this invariance means is that when using a distance that is invariant under the removal of empty bins comparisons within a dataset do not change when the data is modified by Algorithm A.2. The fact that comparisons within a dataset remain the same when the data is modified in order to compare it with other datasets is a property of binning not shared by most data-driven methods.

A.4 The Binary / Summed Binary Data Equivalence

For sufficiently small bin size the binary binned data and the summed binary binned data as produced by Algorithm A.1 become the same. For a given dataset, let $\mathbb{X}^{(binary)}$ be the binary binned data matrix and $\mathbb{X}^{(summed)}$ be the summed binary binned data matrix produced by Algorithm A.1 with some fixed bin size b.

Definition 18. Binary / summed binary data equivalence: *The binary / summed binary data equivalence is said to hold (for a particular dataset) for a given bin size* b *when,*

$$\mathbb{X}^{(binary)} = \mathbb{X}^{(summed)}$$

Using bin sizes for which either the binary / summed binary data equivalence holds, is important in the context of MALDI-MSI data for a number of reasons, including:

- When the binary / summed binary data equivalence holds, there is no spectra with more than one peak in any bin. Having multiple peaks in a single bin can confuse interpretations as, at least in principle (disregarding measurement errors, which are small), each *m/z* value should correspond to a different molecular species and as such it does not make sense to treat them as 'the same' at this level.

- The point above allows for the unambiguous use of non-binary values by substituting these values for the non-zero entries in the binned matrix produced by Algorithm A.1. We discuss this in more detail below. We use this to compare the use of these non-binary data types to the binary data in Section 2.3.3, and consider non-binary data types for classification in Chapter 5.

Let $m_{(1)j}, m_{(2)j}, \ldots, m_{(N_j)j}$ be the sorted (increasing) *m/z* locations of the N_j peaks in the j^{th} spectrum of a given dataset in peaklist form.

Definition 19. Bound on the binary / summed binary data equivalence: *The binary / summed binary data equivalence (Definition 18) holds for all bin sizes $b < b^*$ where*

$$b^* = \min_{j,\, i \in [2, N_j]} \left\{ m_{(i)j} - m_{(i-1)j} \right\}$$

When the binary / summed binary data equivalence holds, a bijection exists between peaks in the dataset and non-zero entries of $\mathbb{X} = \mathbb{X}^{(binary)} = \mathbb{X}^{(summed)}$. This bijection allows us to replace the non-zero entries of $\mathbb{X}$ with some other measure of the presence of the peak they are associated to without any ambiguity as to how this should be done. Up to now we have only considered the binary "peak exists, peak does not exist" indicator for peak presence. Some properties that could be used as a measure of peak presence include:

- Intensity: the maximum height of the peak.

- Area: the integrated area under the peak.

- SNR: the Signal-to-Noise Ratio for the peak.

In Section A.1 we considered only the *m/z* location of peaks for Algorithm A.1, and discarded the other properties recorded on each peak. The binary / summed binary data equivalence (Definition 18) provides us with a method to consider the other peak properties discarded in Section A.1 in a systematic manner. We further explore the idea underpinning Definition 19 in Section 2.3.1, where we show how these ideas can be used to identify an appropriate range for the bin size parameter b used in the binning (Algorithm A.1). In Section 2.3.3 and Section 4.2 we then make use of Definition 18 in order to consider alternative indicators for peak presence such as intensity, area, and SNR in the ovarian and endometrial cancer datasets respectively.

A.5 Binning with Shifted Bin Locations

As mentioned briefly in Section A.4, in some circumstances binning can be sensitive to choice of bin locations. Algorithm A.3 is a modification of Algorithm A.1 that produces binned data with bin centres shifted by some constant c $\left(\frac{-b}{2} \leq c \leq \frac{b}{2}\right)$ relative to those produced by Algorithm A.1.

Algorithm A.3. Binning with shifted bins: *For a given bin size $b > 0$ and c $\left(\frac{-b}{2} \leq c \leq \frac{b}{2}\right)$ follow Algorithm A.1 except replace step 1. with*

1. *Construct $(n_{last} - n_1)$ intervals (bins) of width b and with centres*

$$n_{first}b + c,\ (n_{first} + 1)b + c,\ (n_{first} + 2)b + c,\ \ldots,\ n_{last}b + c$$

(left endpoint open, right closed)

In Section 2.6.2 we consider combined two shifted-bin analyses in order to ensure we do not miss any features of interest due to binning artefacts. Similarly in Section 4.2.1, we try to leverage all the information in the endometrial cancer data by using shifted-bin analyses in parallel to construct meta-classification rules based on a majority of the shifted-bin analyses, thereby addressing any sensitivity the classification may have to choice of bin locations.

Appendix B

Detailed Consideration of Ovarian Datasets

In Section 3.1 we considered the application of DIPPS-based feature extraction as introduced in Chapter 2 to the ovarian cancer data discussed in Section 1.5.1. This feature extraction method yields a set of DIPPS-features that are good positive indicators for a subset of interest in the data. We use clustering to identify clusters subsets of the data roughly corresponding to tissue types, and we compare the sets of DIPPS-features extracted on the basis of these clusters to investigate within and between patient variability, ultimately demonstrating that within patient variability can be sufficiently compensated for in order to detect between patient differences in these data. In the process of investigating these data in Section 3.1 we focus on parts of the data that we are particularly interested in — namely between-patient comparison of cancerous tumour regions. Due to this focus, we omitted details of the comparisons within patients B and C, and comparisons of non-cancer regions between patients. Here we include these omitted results, including brief discussion for completeness.

B.1 Detailed Jaccard Comparisons in Patient B

Figure B.1 shows the Jaccard distance based comparisons analogous to those of Figure 3.2, but for the clustering results shown in Figure 3.3. The discussion of the clustering results in Section 3.1.3 leads to some natural consequences in Figure B.1:

- The purple clusters of B2, B3 and B4 show similarity to both the cancer and stroma of B1. This is expected as these purple clusters contain both cancer and stroma tissue regions.

- The green clusters of B2 and B4 show similarity to the off-tissue of B1. This is similarly expected as these green clusters correspond to off-tissue regions. The salmon and green clusters in datasets B2 and B4 show significant similarity, both within their datasets, and across datasets, and this is expected for the same reason — they correspond to off-tissue regions.

- The green cluster of B3 shows similarity to the cancer cluster of B1. Again, expected because this green cluster corresponds to cancerous tissue. It is particularly notable that although this green cluster shows similarity to both the cancer and stroma, it shows stronger similarity to the cancer. The individual variables responsible for these similarities could potentially be of interest

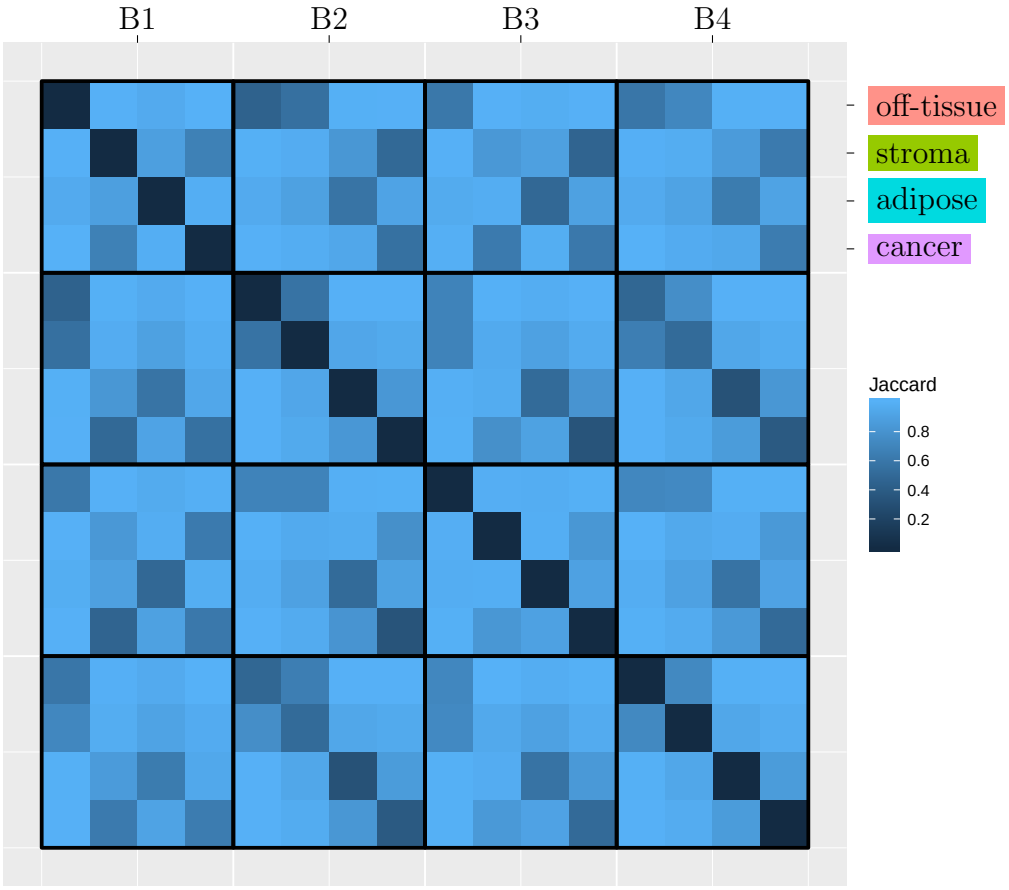


Figure B.1: Image representing the Jaccard distance comparisons of Section 3.1.1 of the cluster memberships of Figure 3.3. A set of DIPPS-features is found for each of the 16 clusters shown in Figure 3.3 using the feature extraction approach discussed in Section 2.6.2 and the heuristic cutoff of Definition 12. The image shown above represents pairwise Jaccard distances between these sets of DIPPS-features. Black lines separate datasets, with the four pixels within each black divisor corresponding to the four clusters for that dataset.

for further investigation into molecular markers of tumour heterogeneity, for example.

- The adipose clusters across all the datasets show good agreement.

Overall, the within-patient comparisons of Figure B.1 agree well with the clustering results of Figure 3.3. If the relationship between the clustering results and the tissue types demonstrate is kept in mind, then the within-patient comparisons of Figure B.1 also exhibit a similar degree of reproducibility to the comparisons of Figure 3.2 between multiple datasets originating from sections of the same tissue block.

B.2 Detailed Jaccard Comparisons in Patient C

Figure B.2 shows the Jaccard distance based comparisons analogous to those of 3.2 and B.1, but for the clustering results shown in Figure 3.4. In Figure B.2, apart from the expected effects of the spreading of the adipose area into the off-tissue area in two of the datasets, strong darkened diagonals are clearly visible. The only

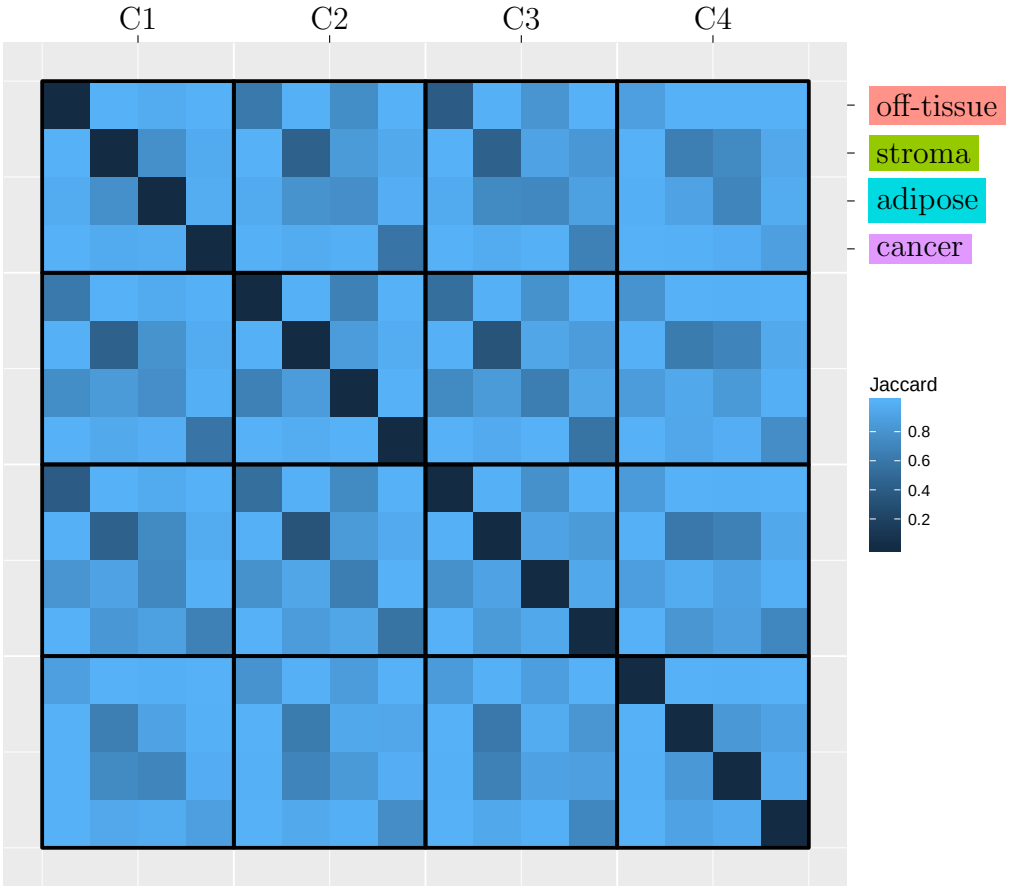


Figure B.2: Image representing the Jaccard distance comparisons of Section 3.1.1 of the cluster memberships of Figure 3.4. A set of DIPPS-features is found for each of the 16 clusters shown in Figure 3.4 using the feature extraction approach discussed in Section 2.6.2 and the heuristic cutoff of Definition 12. The image shown above represents pairwise Jaccard distances between these sets of DIPPS-features. Black lines separate datasets, with the four pixels within each black divisor corresponding to the four clusters for that dataset.

other thing to note in Figure B.2 is that some of the cancer clusters have a notable similarity to the stroma clusters, and this could be due to the purple cluster in some of these datasets (C3, for example) including some surrounding stroma tissue despite primarily corresponding to cancer tissue.

B.3 Between Patient Comparisons

Before I discuss results and interpretations for the between patient comparisons, I provide a quick summary of the conclusions from the within patient comparisons for all three patients:

- **Patient A** (Section 3.1.2): All the clustering results for patient A agreed very well, the only notable deviation being the purple clusters of datasets A1 and A2 included some of the thin connective stroma tissue inbetween adipose regions, which in datasets A3 and A4 are largely grouped into their adipose clusters. It was also noted that the adipose and stroma clusters where the least well separated of the four, possibly because of this connective region that

is clustered with the cancer in A1 and A2 consisting of partly stroma tissue, but being grouped with the adipose in clusters A3 and A4.

- **Patient B** (Section 3.1.3): The clustering results for B1 separated four clusters that corresponded to tissue types reasonably well. In the other three datasets however, the green cluster did not correspond well to any one tissue type, in B3 only corresponding to a small part of stroma tissue, and in B2 and B4 corresponding to off-tissue regions. The remaining stroma regions in A2, A3 and A4 being included in the purple cluster for these three datasets.

- **Patient C** (Section 3.1.4): The data for C2 and C4 showed a degree of speckling, suggesting the possibility of lower quality data. The clustering results for all four datasets corresponded well to tissue types, with the only notable exception being that in C2 and C3 the cyan cluster extended somewhat beyond the tissue and into the off-tissue region — possibly suggesting delocalisation of some peptides.

Figure B.3 shows the Jaccard distance based comparisons between the four clusters in each of the twelve datasets introduced so far. This constitutes 1176 pairwise comparisons and makes Figure B.3 difficult to interpret due to the large amount of information it contains. To aid in the interpretation of Figure B.3 we break it up into smaller block matrices. The three submatrices separated by black dividing lines constituting a block diagonal in Figure B.3 are the same matrices shown in Figure 3.2, Figure B.1 and Figure B.2 respectively — describing the comparisons within patients A, B, and C respectively. We are interested in the between patients comparisons, contained in the triangular block matrix above (or equivalently below, as the matrix is symmetric) the aforementioned block diagonal. These between patient comparisons are split into 3 blocks, corresponding to pairwise comparisons between the three patients:

- Patient A versus Patient B in the centre top (or left centre),

- Patient A versus Patient C in the right top (or left bottom) and

- Patient B versus Patient C in the right centre (or centre bottom)

of Figure B.3. I will consider each of these pairwise comparison blocks individually.

Patient A vs. patient B

- Off-tissue areas agree well, which is not surprising and only notable because the green clusters of B2 and B4 that also occur in off-tissue areas agree with this similarity — supporting the hypothesis that the green clusters extended into the off-tissue region is only a minor phenomena and not a significant effect, as these regions are very similar to the off-tissue regions of not only the other patient B sections but also to the off-tissue regions of the patient A sections.

- Adipose clusters agree well overall. The cyan clusters of patient B show a notable similarity to the stroma clusters of patient A — this likely reflects that they contain some stroma tissue, perhaps due to the stroma of patient B being difficult to differentiate from the adipose. This is further supported by the fact that B1 (where the adipose cluster does not include stroma) does not show this similarity to the stroma clusters of the patient A datasets, and similarly B4 (whose adipose cluster contains the least stroma of B2, B3, and

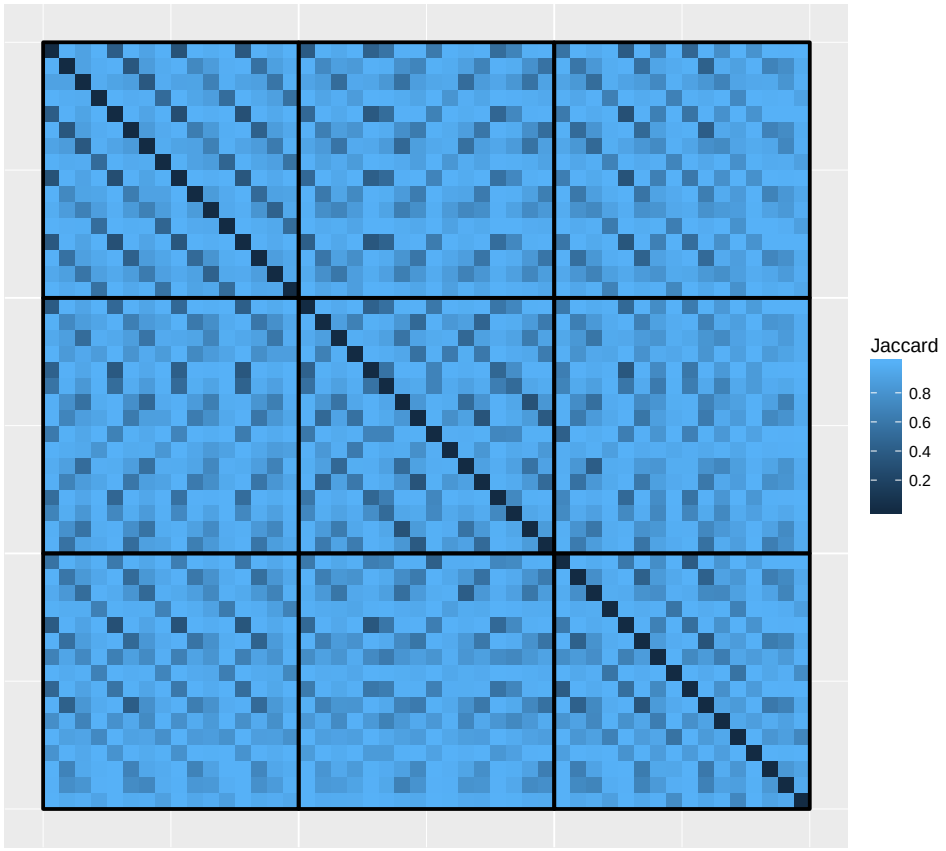


Figure B.3: Image representing the Jaccard distance comparisons of Section 3.1.1 of the cluster memberships of Figures 3.1, 3.3, and 3.4. Black lines separate patients.

B4 due to its cancer cluster including more stroma than B2 or B3) shows much less similarity to the stroma clusters of the patient A datasets than B2 or B3.

- The similarity between cancer clusters is visible, although notably weak. The cancer clusters of patient B additionally show a notable similarity to the stroma clusters of patient A (stronger than their similarity to the cancer clusters of patient A), most likely because the cancer clusters in patient B are not well separated and contain a significant amount of stroma tissue, but nonetheless this demonstrates that cross-patient similarities in stroma tissue can be detected in this way. This is supported by the fact this similarity between cancer of patient B and stroma of patient A is much weaker in dataset B1, where the cancer and stroma are better separated. Also of interest is the similarity between the stroma cluster of dataset B3 and the cancer clusters of patient A — indicating perhaps this small region is not just stroma as was believed, but shares many molecular markers of not only the cancer of patient B, but also of patient A.

Patient A vs. patient C

- Overall good similarity across all clusters, the only broad exception being the notable similarity between adipose and stroma clusters in both directions — possibly simply due to the difficulty to separating these tissue types mentioned previously.

- The spreading of the cyan clusters into off-tissue regions in C2 and C3 is also notable by the similarity of the cyan clusters in these datasets to the off-tissue in patient A.

Patient B vs. patient C

- The cancer clusters of patient B show strong similarity to the stroma of the patient C datasets, most likely due to how the cancer clusters of patient B tend to include a significant amount of surrounding stroma tissue. This similarity is least in B1, whose cancer cluster is best separated from the stroma of the patient B datasets.

- The adipose clusters of datasets C2 and C3 show a similarity to the off-tissue clusters of the patient B datasets, which is expected as the cyan clusters in datasets C2 and C3 extend into off-tissue regions. Otherwise, adipose clusters match up remarkably well.

- Similarly, and unsurprisingly, the stroma clusters of datasets B2 and B4 (which extend into off-tissue regions) show strong similarity to the off-tissue clusters of the patient C datasets.

Appendix C

Matrix Inverse

Here we provide a derivation of the analytic form for the inverse of a class of matrices $A(a, b, c, d)$ defined in Equation C.1. First we introduce notation, shorthand, and preliminary results in Section C.1. We also use many common linear algebra results throughout, and references for these can be found in Halmos (1958); Brookes (2011).

C.1 Notation and Preliminary Results

C.1.1 $A(a, b, c, d)$

We will be interested in matrices of the form

$$A(a,b,c,d) = \begin{bmatrix} a\mathrm{I}_{b\times b} & \mathbf{1}_{b\times d} \\ \mathbf{1}_{b\times d} & c\mathrm{I}_{d\times d} \end{bmatrix} = \begin{bmatrix} a & 0 & 0 & 0 & 1 & 1 & \dots & 1 \\ 0 & a & 0 & 0 & 1 & 1 & \dots & 1 \\ \vdots & & \ddots & \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & a & 1 & 1 & \dots & 1 \\ 1 & 1 & 1 & 1 & c & 0 & \dots & 0 \\ 1 & 1 & 1 & 1 & 0 & c & \dots & 0 \\ \vdots & & & \vdots & \vdots & & \ddots & \vdots \\ 1 & 1 & 1 & 1 & 0 & 0 & \dots & c \end{bmatrix}. \tag{C.1}$$

We will be dealing with matrices of minors of such matrices, and as such some shorthand notation will be useful:

- Let $A \setminus (i, j)$ denote the submatrix of A constructed by removing the ith row and jth column.
- Let $A \setminus (i, .)$ denote the submatrix of A constructed by removing the ith row, and $A \setminus (., j)$ denote the submatrix of A constructed by removing the jth column.
- Let $A \setminus \{(i, j), (k, l)\}$ denote the submatrix of A constructed by removing the ith and kth rows, and jth and lth columns.

Note that $A \setminus \{(i, j), (k, l)\} = A \setminus \{(i, l), (k, j)\} = A \setminus \{(k, l), (i, j)\}$. Furthermore if $j > i$ then $A \setminus \{(i, i), (j, j)\} = (A \setminus (j, j)) \setminus (i, i)$.

Also, the shorthand notation:

$$A_n = A(a, b - n, c, d - n), \qquad n \leq min(b, d) \tag{C.2}$$

will be useful as dropping the (a, b, c, d) dependence is convenient when these values are constant, as they will be in the cases we consider.

C.1.2 Preliminary Results for $|A(a,b,c,d) \setminus (i,j)|$

Here we provide expansions of $|A(a,b,c,d) \setminus (i,j)|$ for all possible (i,j). These preliminary results will be useful for the derivation in Section C.2.

Case 1: $i \in [1,b]$ **and** $j \in [b+1, b+d]$

If $i \neq b$,

$$|A(a,b,c,d) \setminus (i,j)| = -|A(a,b,c,d) \setminus (i+1,j)|$$

If $j \neq b+d$ ($A(a,b,c,d) \setminus (i,j+1)$ makes no sense),

$$|A(a,b,c,d) \setminus (i,j)| = -|A(a,b,c,d) \setminus (i,j+1)|$$

By symmetry these results also hold if i and j are swapped.

Case 2: $i,j \in [1,b]$ **and** $i \neq j$

If $i \neq b$ and $i \neq j-1$,

$$|A(a,b,c,d) \setminus (i,j)| = -|A(a,b,c,d) \setminus (i+1,j)|$$

If $j \neq b$ and $j \neq i-1$,

$$|A(a,b,c,d) \setminus (i,j)| = -|A(a,b,c,d) \setminus (i,j+1)|$$

Case 3: $i,j \in [b+1, b+d]$ **and** $i \neq j$

If $i \neq b+d$ and $i \neq j-1$,

$$|A(a,b,c,d) \setminus (i,j)| = -|A(a,b,c,d) \setminus (i+1,j)|$$

If $j \neq b+d$ and $j \neq i-1$,

$$|A(a,b,c,d) \setminus (i,j)| = -|A(a,b,c,d) \setminus (i,j+1)|$$

C.2 Inverse of $A(a,b,c,d)$

Here we present the derivation for the inverse of a matrix of the form $A(a,b,c,d)$ as introduced in Section C.1. This derivation is structured by following these steps:

- We find the determinant of a general matrix $A(a,b,c,d)$ in Section C.2.1
- We find the matrix of minors in Section C.2.2. This is done by the fact that the (i,j)th entry of the matrix of minors is the determinant of the matrix with the ith row and jth column removed.
- The cofactor matrix can be found from the matrix of minors, M as $C_{ij} = (-1)^{i+j} M_{ij}$ where the subscript ij indicates the $(i,j)^{th}$ element of the matrix.
- The adjoint is the transpose of the cofactor matrix $adj(A) = C^T$. $A(a,b,c,d)$ is symmetric, and as such $adj(A) = C = C^T$.
- Finally, we combine the previous results to find the inverse in Section C.2.3. This is done because for an invertible matrix A the inverse can be found from the adjoint of A, $adj(A)$, as $A^{-1} = \frac{1}{|A|} adj(A)$.

C.2.1 Determinant

Here we derive a closed form for the determinant of a matrix of the form $A(a, b, c, d)$. Remember the shorthand $A_n = A(a, b-n, c, d-n)$ for $n \leq min(b, d)$, which means $A(a, b, c, d) = A_0$. Expanding the determinant along the first row gives:

$$
\begin{aligned}
|A_0| &= a|A(a, b-1, c, d)| + \sum_{i=b+1}^{b+d} (-1)^{i+1} |A_0 \setminus (1, i)| \\
&= a|A(a, b-1, c, d)| + (-1)^b d|A_0 \setminus (1, b+1)| && \text{(Sec. C.1.2)} \\
&= a|A(a, b-1, c, d)| + (-1)^b d \sum_{i=1}^{b} (-1)^{b+i} |A_0 \setminus \{(1, b+1), (b+1, i)\}| && \text{(expand } b\text{th row)} \\
&= a|A(a, b-1, c, d)| + d \sum_{i=1}^{b} (-1)^b (-1)^{b+i} |A_0 \setminus \{(b+1, b+1), (1, i)\}| \\
&= a|A(a, b-1, c, d)| + d \sum_{i=1}^{b} (-1)^i |A(a, b, c, d-1) \setminus (1, i)| \\
&= a|A(a, b-1, c, d)| - d|A_1| + d(b-1)|A(a, b, c, d-1) \setminus (1, 2)| && \text{(Sec. C.1.2)} \\
&= a|A(a, b-1, c, d)| - d|A_1| \\
&\quad + d(b-1) \sum_{i=b+1}^{b+d-1} (-1)^i |A(a, b, c, d-1) \setminus \{(1, 2), (2, i)\}| \\
&= a|A(a, b-1, c, d)| - d|A_1| + d(b-1) \sum_{i=b+1}^{b+d-1} (-1)^i |A_1 \setminus (1, i-1)| \\
&= a|A(a, b-1, c, d)| - d|A_1| + d(d-1)(b-1)(-1)^{b+1} |A_1 \setminus (1, b)| && \text{(Sec. C.1.2)} \\
&= a|A(a, b-1, c, d)| - d|A_1| \\
&\quad + d(d-1)(b-1)(-1)^{b+1} \sum_{i=1}^{b-1} (-1)^{i+b-1} |A_1 \setminus \{(1, b), (b, i)\}| \\
&= a|A(a, b-1, c, d)| - d|A_1| \\
&\quad + d(d-1)(b-1) \sum_{i=1}^{b-1} (-1)^i |A(a, b-1, c, d-2) \setminus (1, i)| \\
&= a|A(a, b-1, c, d)| - d|A_1| - d(d-1)(b-1)|A_2| \\
&\quad + d(d-1)(b-1) \sum_{i=2}^{b-1} (-1)^i |A(a, b-1, c, d-2) \setminus (1, i)| \\
&= a|A(a, b-1, c, d)| - d|A_1| - d(d-1)(b-1)|A_2| \\
&\quad + d(d-1)(b-1)(b-2)|A(a, b-1, c, d-2) \setminus (1, 2)| && \text{(Sec. C.1.2)}
\end{aligned}
$$

Notice that this derived result is a recursive relation. If we apply this recursion l times, the lth blue line will be

$$
\begin{aligned}
|A_0| = a|A(a, b-1, c, d)| &- d \sum_{k=1}^{l} \frac{(d-1)!(b-1)!}{(d-k)!(b-k)!} |A_k| \\
&+ \frac{d!(b-1)!}{(d-l)!(b-l-1)!} |A(a, b-l+1, c, d-l) \setminus (1, 2)| \\
&\dots
\end{aligned}
$$

Consider the application of this recursion $b-1$ times — i.e. the $(b-1)$th blue line would be

$$\begin{aligned}|A_0| &= a|A(a,b-1,c,d)| - d\sum_{k=1}^{b-1}\frac{(d-1)!(b-1)!}{(d-k)!(b-k)!}|A_k| + \frac{d!(b-1)!}{(d-b+1)!}|A(a,2,c,d-b+1)\setminus(1,2)| \\ &= a|A(a,b-1,c,d)| - d\sum_{k=1}^{b-1}\frac{(d-1)!(b-1)!}{(d-k)!(b-k)!}|A_k| - \frac{d!(b-1)!}{(d-b)!}c^{d-b}.\end{aligned}$$

Notice that a solution to this recursion relation is

$$|A(a,b,c,d)| = a^{b-1}\left(a - \frac{db}{c}\right)c^d = a^bc^d - bda^{b-1}c^{d-1}. \tag{C.3}$$

To validate, we substitute:

$$\begin{aligned}d\sum_{k=1}^{b-1}\frac{(d-1)!(b-1)!}{(d-k)!(b-k)!}|A_k| &= d\sum_{k=1}^{b-1}\frac{(d-1)!(b-1)!}{(d-k)!(b-k)!}(a^{b-k}c^{d-k} - (b-k)(d-k)a^{b-k-1}c^{d-k-1}) \\ &= d(d-1)!(b-1)!\left[\sum_{k=1}^{b-1}\frac{a^{b-k}c^{d-k}}{(d-k)!(b-k)!} - \sum_{k=1}^{b-1}\frac{a^{b-k-1}c^{d-k-1}}{(d-k-1)!(b-k-1)!}\right] \\ &= d(d-1)!(b-1)!\left[\sum_{k=1}^{b-1}\frac{a^{b-k}c^{d-k}}{(d-k)!(b-k)!} - \sum_{k=2}^{b}\frac{a^{b-k}c^{d-k}}{(d-k)!(b-k)!}\right] \\ &= d(d-1)!(b-1)!\left[\frac{a^{b-1}c^{d-1}}{(d-1)!(b-1)!} - \frac{c^{d-b}}{(d-b)!}\right] \\ &= da^{b-1}c^{d-1} - \frac{d!(b-1)!}{(d-b)!}c^{d-b}.\end{aligned}$$

So

$$\begin{aligned}|A_0| &= a|A(a,b-1,c,d)| - d\sum_{k=1}^{b-1}\frac{(d-1)!(b-1)!}{(d-k)!(b-k)!}|A_k| - \frac{d!(b-1)!}{(d-b)!}c^{d-b} \\ &= a|A(a,b-1,c,d)| - da^{b-1}c^{d-1} + \frac{d!(b-1)!}{(d-b)!}c^{d-b} - \frac{d!(b-1)!}{(d-b)!}c^{d-b} \\ &= a\left(a^{b-1}c^d - (b-1)da^{b-2}c^{d-1}\right) - da^{b-1}c^{d-1} \\ &= a^bc^d - (b-1)da^{b-1}c^{d-1} - da^{b-1}c^{d-1} \\ &= a^bc^d - (b-1+1)da^{b-1}c^{d-1} \\ &= a^bc^d - bda^{b-1}c^{d-1} \\ &= |A_0|,\end{aligned}$$

thus validating the solution of Equation C.3.

C.2.2 Matrix of Minors

Here we will derive the matrix of minors for a matrix of the form $A(a,b,c,d)$. We separate this problem into cases, and find the minors for each case separately.

Case 1: $i = j$ **(diagonal entries)**

This case reduces to the general problem of finding $|A(a,b,c,d)|$, which we have already solved in Section C.2.1 — $|A(a,b,c,d)|$ is given in Equation C.3. So,

$$|A(a,b,c,d)\backslash(i,i)| = \begin{cases} |A(a,b-1,c,d)| = a^{b-1}c^d - (b-1)da^{b-2}c^{d-1} & \text{if } i \in [1,b] \\ |A(a,b,c,d-1)| = a^b c^{d-1} - b(d-1)a^{b-1}c^{d-2} & \text{if } i \in [b+1,b+d] \end{cases}$$

Case 2: $i = 1$ **and** $j \in [b+1, b+d]$

From Section C.2.2(Case 1) we have:

$$\begin{aligned}
\sum_{k=2}^{b-1} \frac{(d-1)!(b-2)!}{(d-k)!(b-k)!}|A_k| &= \sum_{k=2}^{b-1} \frac{(d-1)!(b-2)!}{(d-k)!(b-k)!}(a^{b-k}c^{d-k} - (b-k)(d-k)a^{b-k-1}c^{d-k-1}) \\
&= (d-1)!(b-2)!\left[\sum_{k=2}^{b-1} \frac{a^{b-k}c^{d-k}}{(d-k)!(b-k)!} - \sum_{k=2}^{b-1} \frac{a^{b-k-1}c^{d-k-1}}{(d-k-1)!(b-k-1)!}\right] \\
&= (d-1)!(b-2)!\left[\sum_{k=2}^{b-1} \frac{a^{b-k}c^{d-k}}{(d-k)!(b-k)!} - \sum_{k=3}^{b} \frac{a^{b-k}c^{d-k}}{(d-k)!(b-k)!}\right] \\
&= (d-1)!(b-2)!\left[\frac{a^{b-2}c^{d-2}}{(d-2)!(b-2)!} - \frac{c^{d-b}}{(d-b)!}\right] \\
&= (d-1)a^{b-2}c^{d-2} - \frac{(d-1)!(b-2)!}{(d-b)!}c^{d-b}
\end{aligned}$$

and thus also

$$\begin{aligned}
|A(a,b,c,d-1)\setminus(1,2)| &= -\sum_{k=2}^{l} \frac{(d-1)!(b-2)!}{(d-k)!(b-k)!}|A_k| + \frac{(d-1)!(b-2)!}{(d-l)!(b-l-1)!}|A(a,b-l+1,c,d-l)\setminus(1,2)| \\
&= -\sum_{k=2}^{b-1} \frac{(d-1)!(b-2)!}{(d-k)!(b-k)!}|A_k| + \frac{(d-1)!(b-2)!}{(d-b+1)!}|A(a,2,c,d-b+1)\setminus(1,2)| \\
&= -\left[(d-1)a^{b-2}c^{d-2} - \frac{(d-1)!(b-2)!}{(d-b)!}c^{d-b}\right] - \frac{(d-1)!(b-2)!}{(d-b)!}c^{d-b} \\
&= -(d-1)a^{b-2}c^{d-2}
\end{aligned}$$

and so by expanding along the $(j-1)$th row (i.e. the jth row of A_0), we have our result:

$$\begin{aligned}
|A_0\setminus(1,j)| &= \sum_{k=1}^{b}(-1)^{j+k-1}|A_0\setminus\{(1,j),(j,k)\}| \\
&= \sum_{k=1}^{b}(-1)^{j+k-1}|A(a,b,c,d-1)\setminus(1,k)| \\
&= (-1)^j|A_1| + \sum_{k=2}^{b}(-1)^{j+k-1}|A(a,b,c,d-1)\setminus(1,k)| \\
&= (-1)^j|A_1| + (-1)^{j-1}(b-1)|A(a,b,c,d-1)\setminus(1,2)| && \text{(Sec. C.1.2)} \\
&= (-1)^j\left[a^{b-1}c^{d-1} - (b-1)(d-1)a^{b-2}c^{d-2}\right] + (-1)^{j-1}(b-1)\left[-(d-1)a^{b-2}c^{d-2}\right] && \text{(Case 1)} \\
&= (-1)^j\left[a^{b-1}c^{d-1} - (b-1)(d-1)a^{b-2}c^{d-2} + (b-1)(d-1)a^{b-2}c^{d-2}\right] \\
&= (-1)^j a^{b-1}c^{d-1}
\end{aligned}$$

Case 3: $i \in [2, b]$ **and** $j \in [b+1, b+d]$

We can extend the result from Section C.2.2(Case 2) for $|A(a,b,c,d-1) \setminus (1,2)|$ to $|A(a,b,c,d-1) \setminus (1,i)|$ by the results in Section C.1.2, or by the same recursive argument as in Section C.2.2(Case 1), either way we obtain the identity

$$(-1)^i|A(a,b,c,d-1) \setminus (1,2)| = |A(a,b,c,d-1) \setminus (1,i)| = |A(a,b,c,d-1) \setminus (i,1)|.$$

So this case reduces to be the same as that of Section C.2.2(Case 2).

$$\begin{aligned}
|A_0 \setminus (i,j)| &= \sum_{k=1}^{b} (-1)^{j+k-1}|A_0 \setminus \{(i,j),(j,k)\}| \\
&= \sum_{k=1}^{b} (-1)^{j+k-1}|A(a,b,c,d-1) \setminus (i,k)| \\
&= (-1)^{j+i-1}|A_1| + \sum_{k=1,\, k\neq i}^{b} (-1)^{j+k-1}|A(a,b,c,d-1) \setminus (i,k)| \\
&= (-1)^{j+i-1}|A_1| + (-1)^j(b-1)|A(a,b,c,d-1) \setminus (i,1)| \\
&= (-1)^{j+i-1}|A_1| + (-1)^{j+i}(b-1)|A(a,b,c,d-1) \setminus (1,2)| && \text{(Identity)} \\
&= (-1)^{j+i-1}\left[a^{b-1}c^{d-1} - (b-1)(d-1)a^{b-2}c^{d-2}\right] + (-1)^{j+i}(b-1)\left[-(d-1)a^{b-2}c^{d-2}\right] \\
&= (-1)^{j+i-1}\left[a^{b-1}c^{d-1} - (b-1)(d-1)a^{b-2}c^{d-2} + (b-1)(d-1)a^{b-2}c^{d-2}\right] \\
&= (-1)^{j+i-1}a^{b-1}c^{d-1}
\end{aligned}$$

Note that for $i = 1$ this is the same as Section C.2.2(Case 2), so this result generalises the result of Section C.2.2(Case 2).

Case 4: $i < j$ **and** $i, j \in [1, b]$

From Section C.2.2(Case 1), using the result from Section C.2.2(Case 3) we have that

$$\begin{aligned}
(-1)^b|A_0 \setminus (1,b+1)| &= -|A_1| + (b-1)|A(a,b,c,d-1) \setminus (1,2)| \\
&= -\left[a^{b-1}c^{d-1} - (b-1)(d-1)a^{b-2}c^{d-2}\right] + (b-1)\left[-(d-1)a^{b-2}c^{d-2}\right] \\
&= -a^{b-1}c^{d-1}.
\end{aligned}$$

This gives us that

$$(-1)^{b+1}|A(a,b-1,c,d) \setminus (i,b)| = (-1)^i a^{b-2}c^{d-1}.$$

So, similarly to before, expanding along the $(j-1)$th row gives us our solution:

$$\begin{aligned}
|A_0 \setminus (i,j)| &= \sum_{k=b+1}^{b+d} (-1)^{j+k}|A_0 \setminus \{(i,j),(j,k)\}| \\
&= \sum_{k=b+1}^{b+d} (-1)^{j+k}|A(a,b-1,c,d) \setminus (i,k-1)| \\
&= (-1)^{j+b+1}d|A(a,b-1,c,d) \setminus (i,b)| \\
&= (-1)^{j+i}da^{b-2}c^{d-1}.
\end{aligned}$$

Case 5: $i < j$ **and** $i, j \in [b+1, b+d]$

In Section C.2.2(Case 4) we noted that

$$(-1)^b |A_0 \setminus (1, b+1)| = -a^{b-1}c^{d-1}.$$

By the result of Section C.1.2 we know that this means that

$$(-1)^b |A_0 \setminus (1, i)| = (-1)^{b+i} a^{b-1}c^{d-1}$$

for $i \in [b+1, b+d]$.

So, again, we obtain our solution by expanding the $(j-1)$th row,

$$\begin{aligned} |A_0 \setminus (i,j)| &= \sum_{k=1}^{b} (-1)^{j+k-1} |A_0 \setminus \{(i,j),(j,k)\}| \\ &= \sum_{k=1}^{b} (-1)^{j+k-1} |A(a,b,c,d-1) \setminus (i,k)| \\ &= (-1)^{j+i} b a^{b-1} c^{d-2}. \end{aligned}$$

Case 6: $i > j$

Trivially all the other cases generalise by symmetry ($|A| = |A^T|$).

C.2.3 Inverse

In Section C.2.2 we derived the matrix of minors for a general matrix $A(a,b,c,d)$,

$$M_{ij} = \begin{cases} a^{b-1}c^d - d(b-1)a^{b-2}c^{d-1} & i,j \in [1,b] & i = j \\ a^b c^{d-1} - b(d-1)a^{b-1}c^{d-2} & i,j \in [b+1,b+d] & i = j \\ (-1)^{j+i} d a^{b-2}c^{d-1} & i,j \in [1,b] & i \neq j \\ (-1)^{j+i} b a^{b-1}c^{d-2} & i,j \in [b+1,b+d] & i \neq j \\ (-1)^{j+i-1} a^{b-1}c^{d-1} & i \in [1,b],\ j \in [b+1,b+d] & \text{or } j \in [1,b],\ i \in [b+1,b+d] \end{cases}$$

The cofactors ($C_{ij} = (-1)^{i+j} M_{ij}$) are thus

$$C_{ij} = \begin{cases} a^{b-1}c^d - d(b-1)a^{b-2}c^{d-1} & i,j \in [1,b] & i = j \\ a^b c^{d-1} - b(d-1)a^{b-1}c^{d-2} & i,j \in [b+1,b+d] & i = j \\ d a^{b-2}c^{d-1} & i,j \in [1,b] & i \neq j \\ b a^{b-1}c^{d-2} & i,j \in [b+1,b+d] & i \neq j \\ -a^{b-1}c^{d-1} & i \in [1,b],\ j \in [b+1,b+d] & \text{or } j \in [1,b],\ i \in [b+1,b+d] \end{cases},$$

and as $|A(a,b,c,d)| = a^b c^d - bda^{b-1}c^{d-1} = a^{b-1}c^{d-1}(ac-bd)$ (Equation C.3), we can conclude that the inverse is

$$\left[A_0^{-1}\right]_{ij} = \left[A(a,b,c,d)^{-1}\right]_{ij} = \begin{cases} \frac{ac-bd+d}{a(ac-bd)} & i,j \in [1,b] & i = j \\ \frac{ac-bd+b}{c(ac-bd)} & i,j \in [b+1,b+d] & i = j \\ \frac{d}{a(ac-bd)} & i,j \in [1,b] & i \neq j \\ \frac{b}{c(ac-bd)} & i,j \in [b+1,b+d] & i \neq j \\ \frac{-1}{ac-bd} & \begin{matrix} i \in [1,b],\ j \in [b+1,b+d] \\ j \in [1,b],\ i \in [b+1,b+d] \end{matrix} & \text{or} \end{cases}$$

Appendix D

Classification Results for Vulvar Cancer Data

In Chapter 5 we consider various approaches to the classification of MALDI-MSI TMA data. Specifically, we investigate the effect variations in pre-processing and classification method have on LOO misclassification in the endometrial cancer data of Section 1.5.3. Ultimately, we concluded that CCA variable ranking, and in particular CCA-LDA, outperformed PCA dimension reduction and that choice of dimension reduction approach was the factor having the biggest consistant effect on classification perfomance. Secondarily, we also demonstrated that choice of data type also had an consistant effect on classification performance – the choice to use log-intensity data correlating with better classification performance. Also of interest was that the choice of classification method seemed to have strong interactions with other factors, such as choice of dimension reduction approach and choice of data type. Specifically, when paired with CCA-based variable selection on the log-intensity data, LDA achieved the best classification performance of the methods we considered.

Here we replicate the analysis presented in Chapter 5 on a different dataset – the vulvar cancer dataset also described breifly in Section 1.5.3. Figures D.1–D.6 mirror the results of Figures 5.1–5.6 exactly, relating to the vulvar data instead of the endometrial data.

We discussed and compared these results with the conclusions drawn from the analysis of the endometrial data in Section 5.6.

Figure D.1: **Classification Without Dimension Reduction.** LOO misclassification on the y-axis using DWD or NB vs. data type on the x-axis.

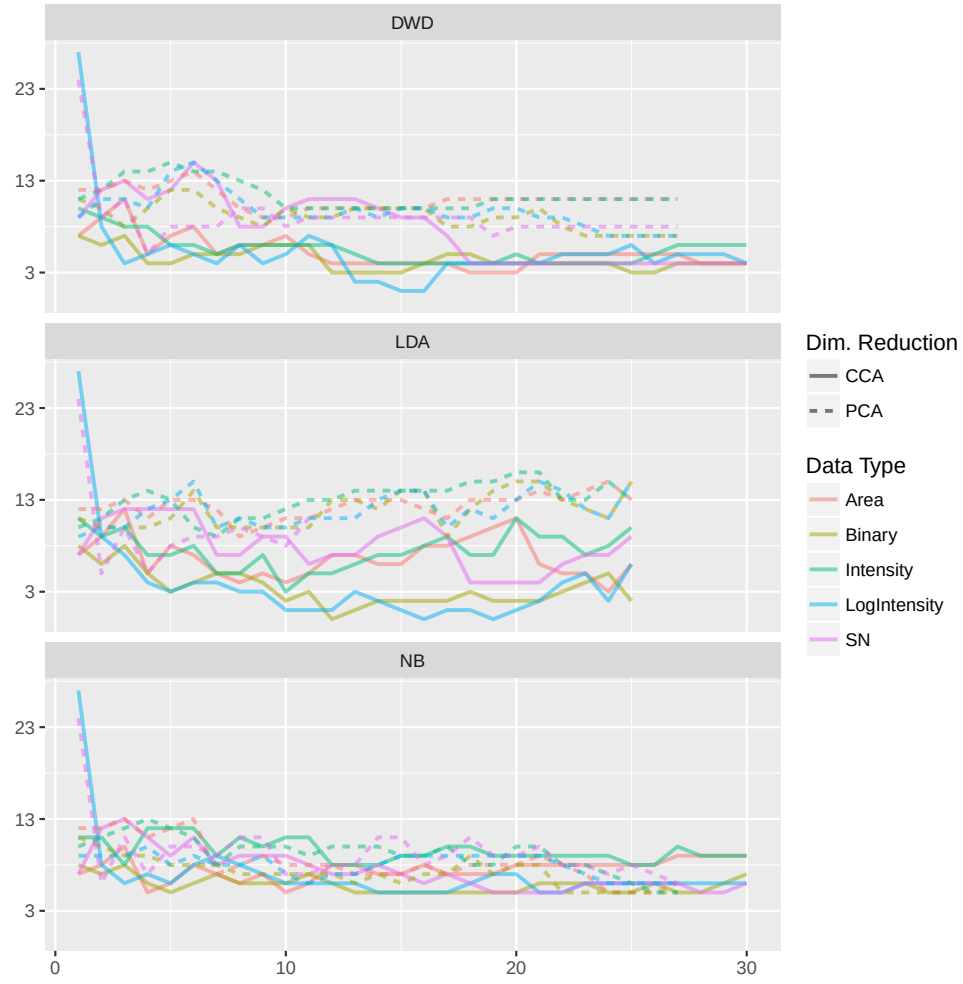


Figure D.2: **Classification of Dimension Reduced Data.** LOO misclassification on the y-axis vs. the number of principal components on the x-axis for the PCA dimension reduced data, or the number of variables retained for the CCA variable reduced data. The results from using each classification method (NB, LDA, and DWD) are shown in seperate panels. Within each panel, results from using each data type are identified by colour. The LOO misclassification refers to the number of incorrectly classified patients out of 28.

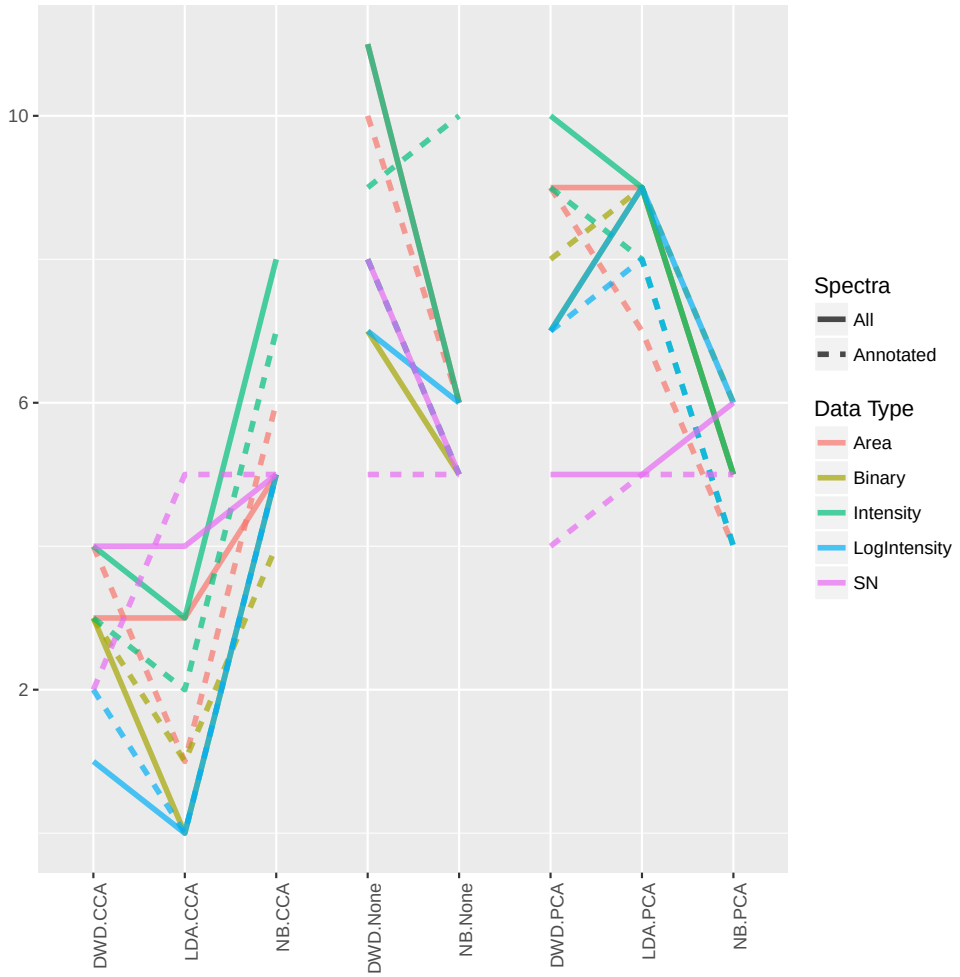


Figure D.3: **Classification With/ Without Restricting to only Cancer Annotated Spectra.** LOO misclassification on the y-axis vs. the combination of classification and dimension reduction method used on the x-axis. The results from using all spectra and using only annotated tumour spectra are identified by use of solid and dashed lines respectively. The results from using each data type are identified by a single colour. In cases that include a dimension reduction step (PCA or CCA), results are only shown for the optimal choice for the number of dimensions, k, that is the k that achieves the lowest LOO misclassification. In cases when there are multiple k that achieve equal lowest LOO misclassification, we choose the smallest — the most parsimonous.

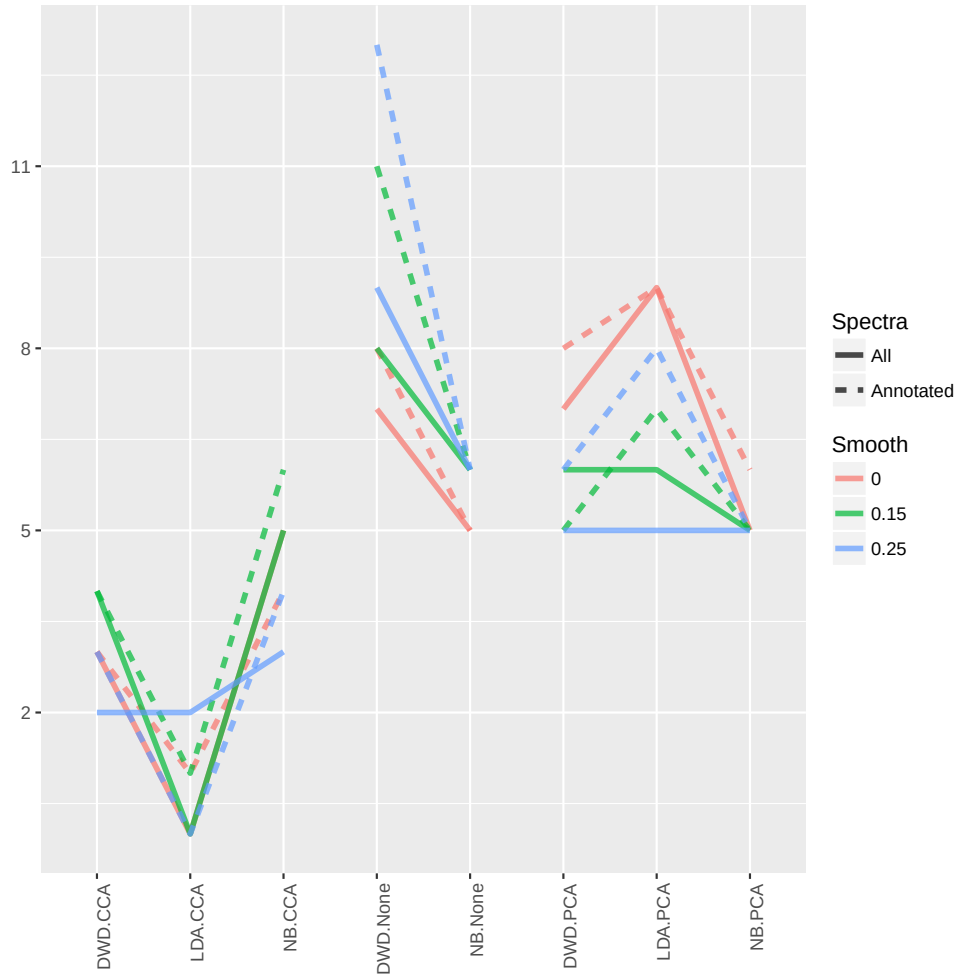


Figure D.4: **Classification of Binary Data With/ Without Spatial Smoothing.** LOO misclassification on the y-axis vs. the combination of classification and dimension reduction method selected on the x-axis. The results from using all spectra and using only annotated tumour spectra are identified by use of solid and dashed lines respectively. The results from using no smoothing ($\tau = 0$), weak smoothing ($\tau = 0.15$), or medium smoothing ($\tau = 0.25$) are identified with colours. The smoothing is described in Section 2.5. In cases that include a dimension reduction step (PCA or CCA), results are only shown for the optimal choice for the number of dimensions, k, that is the k that achieves the lowest LOO misclassification. In cases when there are multiple k that achieve equal lowest LOO misclassification, we choose the smallest — the most parsimonous.

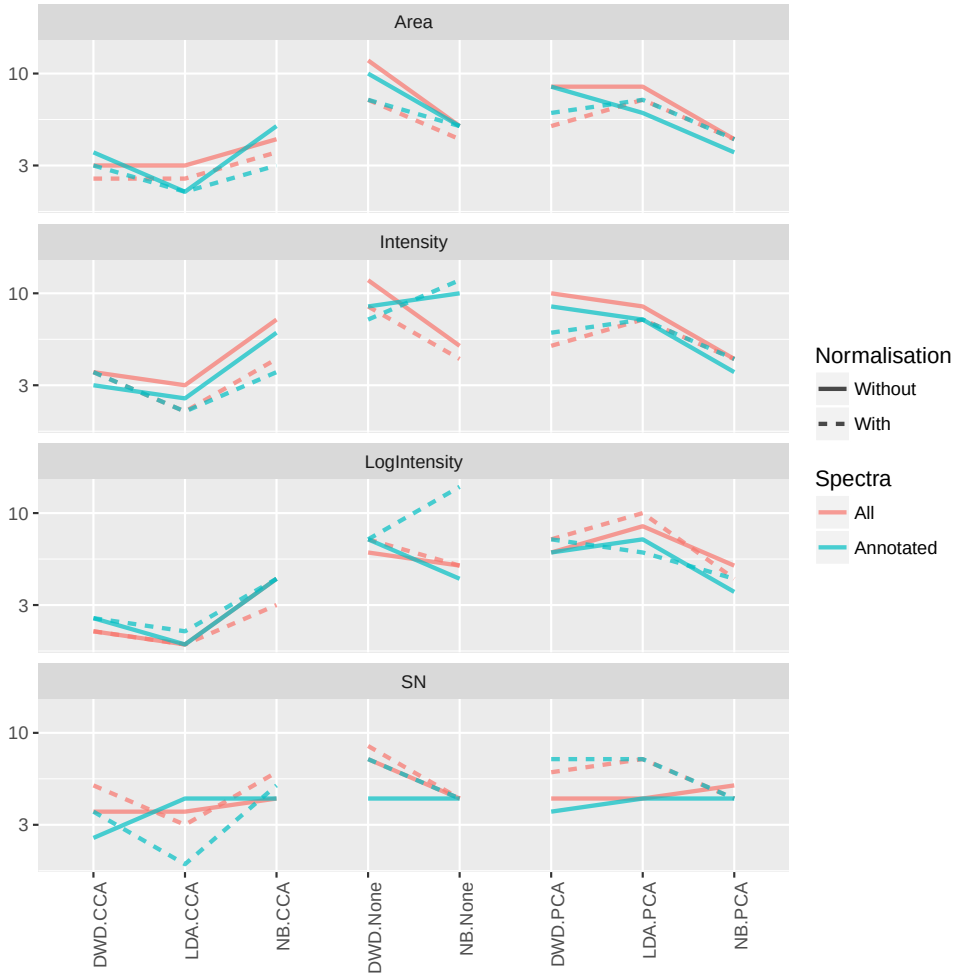


Figure D.5: **Classification of Non-Binary Data With/ Without Normalisation — Part 1: Including Zeroes for Missing Values.** LOO misclassification on the y-axis vs. the combination of classification and dimension reduction method selected on the x-axis. The results from using each data type are shown in seperate panels. The results from using all spectra and using only annotated tumour spectra are identified by two colours respectively accross panels. The results from not using/ using normalisation are identified by use of solid and dashed lines respectively. All results shown include zeros for absent peaks when averaging. In cases that include a dimension reduction step (PCA or CCA), results are only shown for the optimal choice for the number of dimensions, k, that is the k that achieves the lowest LOO misclassification. In cases when there are multiple k that achieve equal lowest LOO misclassification, we choose the smallest — the most parsimonous.

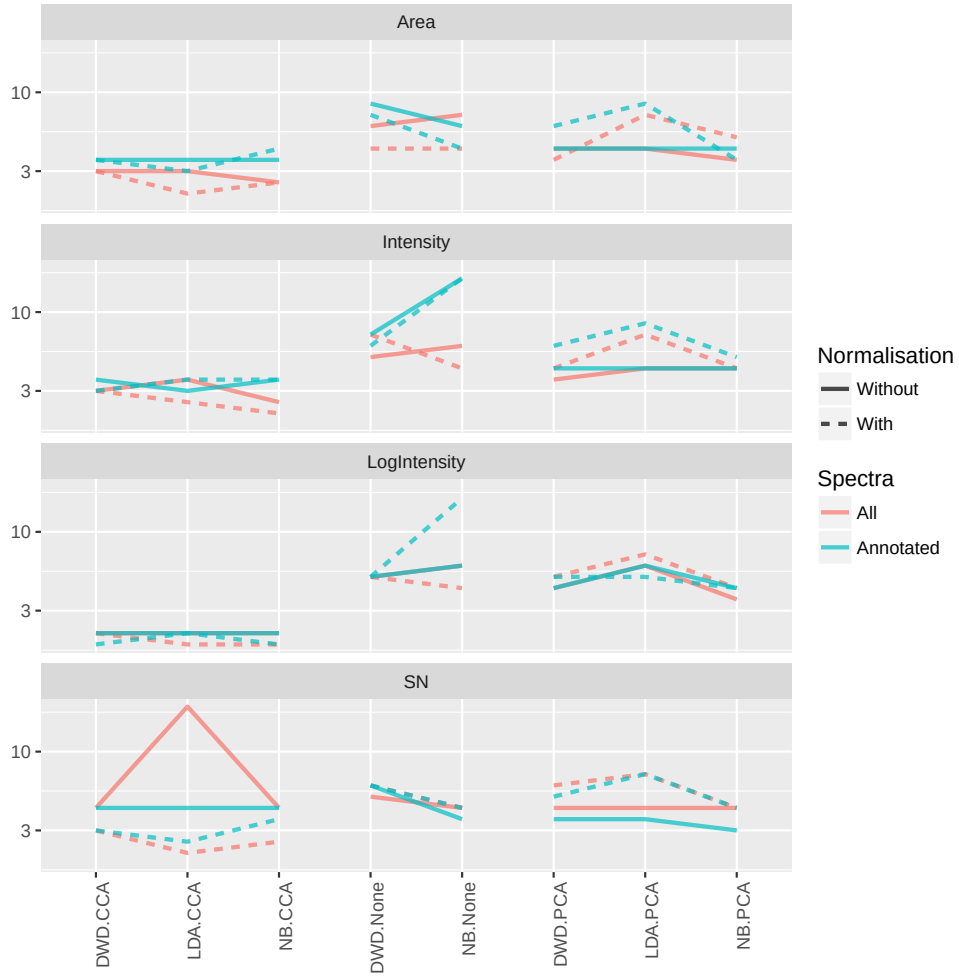


Figure D.6: **Classification of Non-Binary Data With/ Without Normalisation — Part 2: Not Including Missing Values.** LOO misclassification on the y-axis vs. the combination of classification and dimension reduction method selected on the x-axis. The results from using each data type are shown in seperate panels. The results from using all spectra and using only annotated tumour spectra are identified by two colours respectively accross panels. The results from not using/ using normalisation are identified by use of solid and dashed lines respectively. All results shown do not include zeros for absent peaks when averaging. In cases that include a dimension reduction step (PCA or CCA), results are only shown for the optimal choice for the number of dimensions, k, that is the k that achieves the lowest LOO misclassification. In cases when there are multiple k that achieve equal lowest LOO misclassification, we choose the smallest — the most parsimonous.

Bibliography

Gynaecological cancers in australia: an overview, 2012. Cancer series no. 70. Cat. no. CAN 66. Canberra: AIHW.

Karen L. Abbott, Alison V. Nairn, Erica M. Hall, Marc B. Horton, John F. McDonald, Kelley W. Moremen, Daniela M. Dinulescu, and Michael Pierce. Focused glycomic analysis of the n-linked glycan biosynthetic pathway in ovarian cancer. *Proteomics*, 8(16):3210–3220, 2008. ISSN 1615-9861. URL `http://dx.doi.org/10.1002/pmic.200800157`.

Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

Theodore Alexandrov. MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC Bioinformatics*, 13(Suppl 16):S11, 2012. doi:10.1186/1471-2105-13-S16-S11.

Theodore Alexandrov and Jan Hendrik Kobarg. Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. *Bioinformatics*, 27(13):i230–i238, 2011.

Theodore Alexandrov, Michael Becker, Sören-Oliver Deininger, Günther Ernst, Liane Wehder, Markus Grasmair, Ferdinand von Eggeling, Herbert Thiele, and Peter Maass. Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering. *Journal of Proteome Research*, 9(12): 6535–6546, 2010. doi: 10.1021/pr100734z. URL `http://pubs.acs.org/doi/abs/10.1021/pr100734z`.

Theodore Alexandrov, Ilya Chernyavsky, Michael Becker, Ferdinand von Eggeling, and Sergey Nikolenko. Analysis and interpretation of imaging mass spectrometry data by clustering mass-to-charge images according to their spatial similarity. *Analytical chemistry*, 85(23):11189–11195, 2013.

Antoine HP America and Jan HG Cordewener. Comparative LC-MS: A landscape of peaks and valleys. *Proteomics*, 8(4):731–749, 2008.

Edgar Anderson. The irises of the gaspe peninsula. *Bulletin of the American Iris society*, 59:2–5, 1935.

Yutaka Aoki, Atsuhiko Toyama, Takashi Shimada, Tetsuyoshi Sugita, Chikage Aoki, Yukari Umino, Atsushi Suzuki, Daisuke Aoki, Yataro Daigo, Yusuke Nakamura, et al. A novel method for analyzing formalin-fixed paraffin embedded (FFPE) tissue sections by mass spectrometry imaging. *Proceedings of the Japan Academy. Series B, Physical and biological sciences*, 83(7):205, 2007.

Ognian K. Asparoukhov and Wojtek J. Krzanowski. A comparison of discriminant procedures for binary variables. *Computational Statistics & Data Analysis*, 38(2):139 – 160, 2001. ISSN 0167-9473. doi: http://dx.doi.org/10.1016/S0167-9473(01)00032-9. URL http://www.sciencedirect.com/science/article/pii/S0167947301000329.

J Sabine Becker, Miroslav Zoriy, Andreas Matusch, Bei Wu, Dagmar Salber, Christoph Palm, and J Susanne Becker. Bioimaging of metals by laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS). *Mass spectrometry reviews*, 29(1):156–175, 2010.

J Susanne Becker, Ryszard Lobinski, and J Sabine Becker. Metal imaging in non-denaturating 2d electrophoresis gels by laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS) for the detection of metalloproteins. *Metallomics*, 1(4):312–316, 2009.

Adi Ben-Israel and Thomas NE Greville. *Generalized inverses: theory and applications*, volume 15. Springer Science & Business Media, 2003.

U. Boesl, R. Weinkauf, and E.W. Schlag. Reflectron time-of-flight mass spectrometry and laser excitation for the analysis of neutrals, ionized molecules and secondary fragments. *International Journal of Mass Spectrometry and Ion Processes*, 112 (23):121 – 166, 1992. ISSN 0168-1176. doi: 10.1016/0168-1176(92)80001-H. URL http://www.sciencedirect.com/science/article/pii/016811769280001H.

David Bonnel, Rmi Longuespee, Julien Franck, Morad Roudbaraki, Pierre Gosset, Robert Day, Michel Salzet, and Isabelle 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*, 401:149–165, 2011. ISSN 1618-2642. URL http://dx.doi.org/10.1007/s00216-011-5020-5.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. ISSN 0885-6125. URL http://dx.doi.org/10.1023/A:1010933404324.

Mike Brookes. The matrix reference manual, 2011. URL http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/intro.html.

Corrie Brown. Antigen retrieval methods for immunohistochemistry. *Toxicologic pathology*, 26(6):830–831, 1998.

Ricardo JGB Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172. Springer, 2013.

Rita Casadonte and Richard M Caprioli. Proteomic analysis of formalin-fixed paraffin-embedded tissue by MALDI imaging mass spectrometry. *Nature protocols*, 6(11):1695–1709, 2011.

George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Pierre Chaurand, Melinda E. Sanders, Roy A. Jensen, and Richard M. Caprioli. Proteomics in diagnostic pathology: Profiling and imaging proteins directly in tissue sections. *The American Journal of Pathology*, 165(4):1057 – 1068, 2004.

ISSN 0002-9440. doi: http://dx.doi.org/10.1016/S0002-9440(10)63367-6. URL http://www.sciencedirect.com/science/article/pii/S0002944010633676.

Mo Chen, Jian Zhang, and James L Manley. Turning on a fuel switch of cancer: hnrnp proteins regulate alternative splicing of pyruvate kinase mrna. *Cancer research*, 70(22):8977–8980, 2010.

John Conway. The game of life. *Scientific American*, 223(4):4, 1970.

Dale S Cornett, Michelle L Reyzer, Pierre Chaurand, and Richard M Caprioli. MALDI imaging mass spectrometry: molecular snapshots of biochemical systems. *Nature methods*, 4(10):828–833, 2007.

David R Cox. The analysis of multivariate binary data. *Journal of the Royal Statistical Society. Series C (Applied Statistics).*, 21(2):113–120, 1972.

WT Creasman, F Odicino, P Maisonneuve, MA Quinn, U Beller, JL Benedet, APM Heintz, HYS Ngan, and S Pecorelli. Carcinoma of the corpus uteri. *International Journal of Gynecology & Obstetrics*, 95:S105–S143, 2006.

Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

David K Crockett, Zhaosheng Lin, Cecily P Vaughn, Megan S Lim, and Kojo SJ Elenitoba-Johnson. Identification of proteins from formalin-fixed paraffin-embedded cells by LC-MS/MS. *Laboratory investigation*, 85(11):1405–1415, 2005.

Valerie V Cross and Thomas A Sudkamp. *Similarity and compatibility in fuzzy set theory: assessment and applications*, volume 93. Springer Science & Business Media, 2002.

Sören-Oliver Deininger, Matthias P. Ebert, Arne Futterer, Marc Gerhard, and Christoph Röcken. MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *Journal of Proteome Research*, 7(12):5230–5236, 2008. URL http://pubs.acs.org/doi/abs/10.1021/pr8005777. PMID: 19367705.

Marie-Claude Djidja, Emmanuelle Claude, Marten F Snel, Simona Francese, Peter Scriven, Vikki Carolan, and Malcolm R Clench. Novel molecular tumour classification using maldi–mass spectrometry imaging of tissue micro-array. *Analytical and bioanalytical chemistry*, 397(2):587–601, 2010.

Sandrine Dudoit, Jane Fridlyand, and Terence P Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87, 2002. URL http://amstat.tandfonline.com/doi/abs/10.1198/016214502753479248.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

Brian S Everitt and Anders Skrondal. *The Cambridge dictionary of statistics*. 2002.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861 – 874, 2006. ISSN 0167-8655. doi: http://dx.doi.org/10.1016/j.patrec.2005.10.010. URL http://www.sciencedirect.com/science/article/pii/S016786550500303X. ROC Analysis in Pattern Recognition.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

C.H. Fox, F.B. Johnson, J. Whiting, P.P. Roller, et al. Formaldehyde fixation. *J Histochem Cytochem*, 33(8):845–853, 1985.

Megan M. Gessel, Jeremy L. Norris, and Richard M. Caprioli. MALDI imaging mass spectrometry: Spatial molecular analysis to enable a new age of discovery. *Journal of Proteomics*, 107:71 – 82, 2014. ISSN 1874-3919. doi: http://dx.doi.org/10.1016/j.jprot.2014.03.021. URL `http://www.sciencedirect.com/science/article/pii/S1874391914001432`. Special Issue: "20 years of Proteomics" in memory of Vitaliano Pallini.

Karin Gorzolka and Axel Walch. MALDI mass spectrometry imaging of formalin-fixed paraffin-embedded tissues in clinical research. *Histology and histopathology*, 29(11):1365–1376, November 2014. ISSN 0213-3911. URL `http://europepmc.org/abstract/MED/24838644`.

Lawrence Gray. A mathematician looks at wolfram's new kind of science. *Notices-American Mathematical Society*, 50(2):200–211, 2003.

M Reid Groseclose, Malin Andersson, William M Hardesty, and Richard M Caprioli. Identification of proteins directly from tissue: in situ tryptic digestions coupled with imaging mass spectrometry. *Journal of Mass Spectrometry*, 42(2):254–262, 2007.

M. Reid Groseclose, Pierre P. Massion, Pierre Chaurand, and Richard M. Caprioli. High-throughput proteomic analysis of formalin-fixed paraffin-embedded tissue microarrays using maldi imaging mass spectrometry. *Proteomics*, 8(18):3715–3724, 2008. ISSN 1615-9861. doi: 10.1002/pmic.200800495. URL `http://dx.doi.org/10.1002/pmic.200800495`.

Sabine Guenther, Andreas Römpp, Wolfgang Kummer, and Bernhard Spengler. AP-MALDI imaging of neuropeptides in mouse pituitary gland with 5 μm spatial resolution and high mass accuracy. *International Journal of Mass Spectrometry*, 305(23):228 – 237, 2011. ISSN 1387-3806. doi: http://dx.doi.org/10.1016/j.ijms.2010.11.011. URL `http://www.sciencedirect.com/science/article/pii/S1387380610004380`. Special Issue: In Recognition of Catherine Costello, Recipient of the 2010 Field & Franklin Award.

Johan O. R. Gustafsson, Martin K. Oehler, Shaun R. McColl, and Peter Hoffmann. Citric acid antigen retrieval (CAAR) for tryptic peptide imaging directly on archived formalin-fixed paraffin-embedded tissue. *Journal of Proteome Research*, 9(9):4315–4328, July 2010. URL `http://pubs.acs.org/doi/abs/10.1021/pr9011766`.

Johan O. R. Gustafsson, Martin K. Oehler, Andrew Ruszkiewicz, Shaun R. McColl, and Peter Hoffmann. MALDI imaging mass spectrometry (MALDI-IMS) – application of spatial proteomics for ovarian cancer classification and diagnosis. *International Journal of Molecular Sciences*, 12(1):773–794, January 2011. ISSN 1422-0067. doi: 10.3390/ijms12010773. URL `http://www.mdpi.com/1422-0067/12/1/773/`.

Johan O.R. Gustafsson, James S. Eddes, Stephan Meding, Tomas Koudelka, Martin K. Oehler, Shaun R. McColl, and Peter Hoffmann. Internal calibrants allow high accuracy peptide matching between MALDI imaging MS and LC-MS/MS. *Journal of Proteomics*, 75(16):5093 – 5105, 2012. ISSN 1874-3919. doi: 10.1016/j.jprot.2012.04.054. URL `http://www.sciencedirect.com/science/article/pii/S1874391912003259`. Special Issue: Imaging Mass Spectrometry: A Users Guide to a New Technique for Biological and Biomedical Research.

Ove Johan Ragnar Gustafsson. *Molecular characterization of metastatic ovarian cancer by MALDI imaging mass spectrometry.* PhD thesis, School of Molecular and Biomedical Science, 2012.

Ove JR Gustafsson, Matthew T Briggs, Mark R Condina, Lyron J Winderbaum, Matthias Pelzing, Shaun R McColl, Arun V Everest-Dass, Nicolle H Packer, and Peter Hoffmann. MALDI imaging mass spectrometry of N-linked glycans on formalin-fixed paraffin-embedded murine kidney. *Analytical and bioanalytical chemistry*, 407(8):2127–2139, 2015. URL `http://link.springer.com/article/10.1007/s00216-014-8293-7`.

Steven P Gygi, Garry L Corthals, Yanni Zhang, Yvan Rochon, and Ruedi Aebersold. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proceedings of the National Academy of Sciences*, 97(17):9390–9395, 2000.

Paul R. Halmos. *Finite-Dimensional Vector Spaces*. 0172-6056. Springer-Verlag New York, 1 edition, 1958. doi: 10.1007/978-1-4612-6387-6.

Brian L. Hood, Marlene M. Darfler, Thomas G. Guiel, Bungo Furusato, David A. Lucas, Bradley R. Ringeisen, Isabell A. Sesterhenn, Thomas P. Conrads, Timothy D. Veenstra, and David B. Krizman. Proteomic analysis of formalin-fixed prostate cancer tissue. *Molecular & Cellular Proteomics*, 4(11):1741–1753, November 2005. doi: 10.1074/mcp.M500102-MCP200. URL `http://www.mcponline.org/content/4/11/1741.abstract`.

James Jaccard and Jacob Jacoby. *Theory construction and model-building skills: A practical guide for social scientists*. Guilford Press, 2010.

Suzanne M Jacques, Faisal Qureshi, Adnan Munkarah, and W Dwayne Lawrence. Interinstitutional surgical pathology review in gynecologic oncology I: Cancer in endometrial curettings and biopsies. *International journal of gynecological pathology*, 17(1):36–41, 1998.

Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

Ahmedin Jemal, Freddie Bray, Melissa M. Center, Jacques Ferlay, Elizabeth Ward, and David Forman. Global cancer statistics. *CA: A Cancer Journal for Clinicians*, 61(2):69–90, 2011. ISSN 1542-4863. URL `http://dx.doi.org/10.3322/caac.20107`.

Emrys A Jones, Alexandra van Remoortere, René JM van Zeijl, Pancras CW Hogendoorn, Judith VMG Bovée, André M Deelder, and Liam A McDonnell. Multiple statistical analysis techniques corroborate intratumor heterogeneity in imaging mass spectrometry datasets of myxofibrosarcoma. *PloS one*, 6(9):e24913, 2011.

Emrys A. Jones, Sören-Oliver Deininger, Pancras C.W. Hogendoorn, André M. Deelder, and Liam A. McDonnell. Imaging mass spectrometry statistical analysis. *Journal of Proteomics*, 75(16):4962 – 4989, 2012. ISSN 1874-3919. doi: http://dx.doi.org/10.1016/j.jprot.2012.06.014. URL `http://www.sciencedirect.com/science/article/pii/S1874391912004575`. Special Issue: Imaging Mass Spectrometry: A Users Guide to a New Technique for Biological and Biomedical Research.

R Kaufmann, B Spengler, and F Lützenkirchen. Mass spectrometric sequencing of linear peptides by product-ion analysis in a reflectron time-of-flight mass spectrometer using matrix-assisted laser desorption ionization. *Rapid communications in mass spectrometry*, 7(10):902–910, 1993.

Sheerin Khatib-Shahidi, Malin Andersson, Jennifer L Herman, Todd A Gillespie, and Richard M Caprioli. Direct molecular analysis of whole-body animal tissue sections by imaging maldi mass spectrometry. *Analytical chemistry*, 78(18):6448–6456, 2006.

Inge Koch. *Analysis of Multivariate and High-Dimensional Data*, volume 32 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2013.

Inge Koch and Kanta Naito. Prediction of multivariate responses with a selected number of principal components. *Computational Statistics & Data Analysis*, 54 (7):1791–1807, 2010.

Dong Hyeon Lee, Kwanghoe Chung, Ji-Ae Song, Tae-heon Kim, Haeyoun Kang, Jin Hyong Huh, Sang-geun Jung, Jung Jae Ko, and Hee Jung An. Proteomic identification of paclitaxel-resistance associated hnRNP A2 and GDI 2 proteins in human ovarian cancer cells. *Journal of proteome research*, 9(11):5668–5676, 2010.

Sang-Ho Lee and Chi-Hyuck Jun. Discriminant analysis of binary data following multivariate bernoulli distribution. *Expert Systems with Applications*, 38(6):7795 – 7802, 2011. ISSN 0957-4174. doi: http://dx.doi.org/10.1016/j.eswa.2010.12.126. URL `http://www.sciencedirect.com/science/article/pii/S0957417410014892`.

Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

Loet Leydesdorff. On the normalization and visualization of author co-citation data: Salton's cosine versus the jaccard index. *Journal of the American Society for Information Science and Technology*, 59(1):77–85, 2008.

Lance A Liotta, Mauro Ferrari, and Emanuel Petricoin. Clinical proteomics: written in blood. *Nature*, 425(6961):905–905, 2003.

Josip Lovric. *Introducing proteomics: from concepts to sample separation, mass spectrometry and data analysis*. John Wiley & Sons, 2011.

Parag Mallick and Bernhard Kuster. Proteomics: a pragmatic perspective. *Nat Biotech*, 28(7):695–709, July 2010. ISSN 1087-0156. URL `http://dx.doi.org/10.1038/nbt.1658`.

J.S. Marron, M.J. Todd, and J. Ahn. Distance-weighted discrimination. *American Statistical Association*, 102(480):1267–1271, 2007.

Nadine E Mascini, Gert B Eijkel, Petra ter Brugge, Jos Jonkers, Jelle Wesseling, and Ron MA Heeren. The use of mass spectrometry imaging to predict treatment response of patient-derived xenograft models of triple-negative breast cancer. *Journal of proteome research*, 14(2):1069–1075, 2015.

Geoffrey J McLachlan and Kaye E Basford. Mixture models. inference and applications to clustering. *Statistics: Textbooks and Monographs, New York: Dekker, 1988*, 1, 1988.

Stephan Meding, Karina Martin, Ove JR Gustafsson, James S Eddes, Sandra Hack, Martin K Oehler, and Peter Hoffmann. Tryptic peptide reference data sets for MALDI imaging mass spectrometry on formalin-fixed ovarian cancer tissues. *Journal of proteome research*, 12(1):308–315, 2012.

Melanie Mitchell et al. Computation in cellular automata: A selected review. *Nonstandard Computation*, pages 95–140, 1996.

Parul Mittal, Manuela Klingler-Hoffmann, Georgia Arentz, Lyron Winderbaum, Noor A Lokman, Chao Zhang, Lyndal Anderson, James Scurry, Yee Leung, Colin JR Stewart, et al. Lymph node metastasis of primary endometrial cancers: Associated proteins revealed by MALDI imaging. *Proteomics*, 2016. URL `http://onlinelibrary.wiley.com/doi/10.1002/pmic.201500455/full`.

C Paul Morrow, Brian N Bundy, Robert J Kurman, William T Creasman, Paul Heller, Howard D Homesley, and James E Graham. Relationship between surgical-pathological risk factors and outcome in clinical stage I and II carcinoma of the endometrium: a gynecologic oncology group study. *Gynecologic oncology*, 40(1): 55–65, 1991.

Jeremy L. Norris, Dale S. Cornett, James A. Mobley, Malin Andersson, Erin H. Seeley, Pierre Chaurand, and Richard M. Caprioli. Processing MALDI mass spectra to improve mass spectral direct tissue analysis. *International Journal of Mass Spectrometry*, 260:212 – 221, 2007. ISSN 1387-3806. doi: http://dx.doi.org/10.1016/j.ijms.2006.10.005. URL `http://www.sciencedirect.com/science/article/pii/S1387380606004714`. Imaging Mass Spectrometry Special Issue.

Shao-En Ong and Matthias Mann. Mass spectrometry–based proteomics turns quantitative. *Nature chemical biology*, 1(5):252–262, 2005.

Stacey R Oppenheimer, Deming Mi, Melinda E Sanders, and Richard M Caprioli. Molecular analysis of tumor margins by maldi mass spectrometry in renal carcinoma. *Journal of proteome research*, 9(5):2182–2190, 2010.

Darryl Erik Palmer-Toy, Bryan Krastins, David A Sarracino, Joseph B Nadol, and Saumil N Merchant. Efficient method for the proteomic analysis of fixed and embedded tissues. *Journal of proteome research*, 4(6):2404–2411, 2005.

Sheng Pan, Ru Chen, Ruedi Aebersold, and Teresa A Brentnall. Mass spectrometry based glycoproteomicsfrom a proteomics perspective. *Molecular & Cellular Proteomics*, 10(1):R110–003251, 2011.

David Peel and Geoffrey J McLachlan. Robust mixture modelling using the t distribution. *Statistics and computing*, 10(4):339–348, 2000.

Roger Penrose. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pages 406–413. Cambridge Univ Press, 1955.

Gereon Poschmann, Barbara Sitek, Bence Sipos, Michael Hamacher, Oliver Vonend, Helmut E. Meyer, and K. Sthler. Cell-based proteome analysis: The first stage in the pipeline for biomarker discovery. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1794(9):1309 – 1316, 2009. ISSN 1570-9639. doi: 10.1016/j.bbapap.2009.07.001. URL `http://www.sciencedirect.com/science/article/pii/S1570963909001617`. Current Topics in Proteins and Proteomics.

David Martin Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 2011.

Sandra Rauser, Claudio Marquardt, Benjamin Balluff, Sören-Oliver Deininger, Christian Albers, Eckhard Belau, Ralf Hartmer, Detlev Suckau, Katja Specht, Matthias Philip Ebert, et al. Classification of HER2 receptor status in breast cancer tissues by MALDI imaging mass spectrometry. *Journal of proteome research*, 9(4):1854–1863, 2010. URL `http://pubs.acs.org/doi/abs/10.1021/pr901008d`.

Carmela Ricciardelli and Martin K. Oehler. Diverse molecular pathways in ovarian cancer and their clinical significance. *Maturitas*, 62(3):270 – 275, 2009. ISSN 0378-5122. doi: 10.1016/j.maturitas.2009.01.001. URL `http://www.sciencedirect.com/science/article/pii/S0378512209000024`.

Adelina Rogowska-Wrzesinska, Marie-Catherine Le Bihan, Morten Thaysen-Andersen, and Peter Roepstorff. 2D gels still have a niche in proteomics. *Journal of proteomics*, 88:4–13, 2013.

Bunja Rungruang and Alexander B Olawaiye. Comprehensive surgical staging for endometrial cancer. *Reviews in obstetrics & gynecology*, 5(1):28–34, 2012.

Bernhard Schölkopf and Alex Smola. Support vector machines. *Encyclopedia of Biostatistics*, 1998.

Kristina Schwamborn and Richard M. Caprioli. Molecular imaging by mass spectrometry - looking beyond classical histology. *Nat Rev Cancer*, 10(9):639–646, September 2010. ISSN 1474-175X. URL `http://dx.doi.org/10.1038/nrc2917`.

Sarah A. Schwartz, Michelle L. Reyzer, and Richard M. Caprioli. Direct tissue analysis using matrix-assisted laser desorption/ionization mass spectrometry: practical aspects of sample preparation. *Journal of Mass Spectrometry*, 38(7):699–708, 2003. ISSN 1096-9888. URL `http://dx.doi.org/10.1002/jms.505`.

Erin H. Seeley and Richard M. Caprioli. Maldi imaging mass spectrometry of human tissue: method challenges and clinical perspectives. *Trends in Biotechnology*, 29(3):136 – 143, March 2011. ISSN 0167-7799. doi: 10.1016/j.

tibtech.2010.12.002. URL http://www.sciencedirect.com/science/article/pii/S0167779910002040.

Shan-Rong Shi, MARC E Key, and KRISHAN 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(6):741–748, 1991.

Shan-Rong Shi, Cheng Liu, Jeanette Perez, and Clive R Taylor. Protein-embedding technique: a potential approach to standardization of immunohistochemistry for formalin-fixed, paraffin-embedded tissue sections. *Journal of Histochemistry & Cytochemistry*, 53(9):1167–1170, 2005.

Stefan Steurer, Carina Borkowski, Sinje Odinga, Malte Buchholz, Christina Koop, Hartwig Huland, Michael Becker, Matthias Witt, Dennis Trede, Maryam Omidi, et al. MALDI mass spectrometric imaging based identification of clinically relevant signals in prostate cancer using large-scale tissue microarrays. *International Journal of Cancer*, 133(4):920–928, 2013.

Glenn Stone, David Clifford, Johan OR Gustafsson, Shaun R McColl, and Peter Hoffmann. Visualisation in imaging mass spectrometry using the minimum noise fraction transform. *BMC research notes*, 5(1):419, 2012.

Lynne Thadikkaran, Michèle A Siegenthaler, David Crettaz, Pierre-Alain Queloz, Philippe Schneider, and Jean-Daniel Tissot. Recent advances in blood-related proteomics. *Proteomics*, 5(12):3019–3034, 2005.

Yukiharu Todo, Ritsu Yamamoto, Shinichiro Minobe, Yoshihiro Suzuki, Umazume Takeshi, Makiko Nakatani, Yukiko Aoyagi, Yoko Ohba, Kazuhira Okamoto, and Hidenori Kato. Risk factors for postoperative lower-extremity lymphedema in endometrial cancer survivors who had treatment including lymphadenectomy. *Gynecologic oncology*, 119(1):60–64, 2010.

ML Vestal, P Juhasz, and SA Martin. Delayed extraction matrix-assisted laser desorption time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, 9(11):1044–1050, 1995.

Rui Vitorino, Maria João C Lobo, António J Ferrer-Correira, Joshua R Dubin, Kenneth B Tomer, Pedro M Domingues, and Francisco ML Amado. Identification of human whole saliva protein components using proteomics. *Proteomics*, 4(4): 1109–1115, 2004.

Axel Walch, Sandra Rauser, Sören-Oliver Deininger, and Heinz Höfler. MALDI imaging mass spectrometry for direct tissue analysis: a new frontier for molecular histology. *Histochemistry and Cell Biology*, 130(3):421–434, 2008. ISSN 0948-6143. URL http://dx.doi.org/10.1007/s00418-008-0469-9. 10.1007/s00418-008-0469-9.

M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman and Hall Ltd, London, 1995.

Valerie C Wasinger, Stuart J Cordwell, Anne Cerpa-Poljak, Jun X Yan, Andrew A Gooley, Marc R Wilkins, Mark W Duncan, Ray Harris, Keith L Williams, and Ian Humphery-Smith. Progress with gene-product mapping of the mollicutes: Mycoplasma genitalium. *Electrophoresis*, 16(1):1090–1094, 1995.

Marc R Wilkins, Christian Pasquali, Ron D Appel, Keli Ou, Olivier Golaz, Jean-Charles Sanchez, Jun X Yan, Andrew A Gooley, Graham Hughes, Ian Humphery-Smith, et al. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Nature Biotechnology*, 14(1):61–65, 1996.

Lyron Winderbaum, Cristiana L Ciobanu, Nigel J Cook, Matthew Paul, Andrew Metcalfe, and Sarah Gilbert. Multivariate analysis of an LA-ICP-MS trace element dataset for pyrite. *Mathematical Geosciences*, 44(7):823–842, 2012. URL `http://link.springer.com/article/10.1007/s11004-012-9418-1`.

Lyron Winderbaum, Inge Koch, Parul Mittal, and Peter Hoffmann. Classification of MALDI-MS imaging data of tissue microarrays using canonical correlation analysis-based variable selection. *Proteomics*, 2016. URL `http://onlinelibrary.wiley.com/doi/10.1002/pmic.201500451/full`.

Lyron J Winderbaum, Inge Koch, Ove JR Gustafsson, Stephan Meding, Peter Hoffmann, et al. Feature extraction for proteomics imaging mass spectrometry data. *The Annals of Applied Statistics*, 9(4):1973–1996, 2015. doi: 10.1214/15-AOAS870. URL `http://arxiv.org/abs/1410.1630v2`.

Daniela M Witten and Robert J Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8(1):1–27, 2009.

Stephen Wolfram. Cellular automata as models of complexity. *Nature*, 311(5985): 419–424, 1984.

Baolin Wu, Tom Abbott, David Fishman, Walter McMurray, Gil Mor, Kathryn Stone, David Ward, Kenneth Williams, and Hongyu Zhao. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19(13):1636–1643, 2003. doi: 10.1093/bioinformatics/btg210. URL `http://bioinformatics.oxfordjournals.org/content/19/13/1636.abstract`.

Guang Xu, Xin Liu, Qing Yan Liu, Yanhong Zhou, and Jianjun Li. Improve accuracy and sensibility in glycan structure prediction by matching glycan isotope abundance. *Analytica chimica acta*, 743:80–89, 2012.

Rui Xu, Donald Wunsch, et al. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678, 2005.

J.R. Yates, C.I. Ruse, and A. Nakorchevsky. Proteomics by mass spectrometry: Approaches, advances, and applications. *Annual Review of Biomedical Engineering*, 11:49–79, 2009.