Statistical analysis of proteomic mass spectrometry data for the identification of biomarkers and disease diagnosis

Tyman Stanford

Thesis submitted for the degree of
Doctor of Philosophy in Statistics at
The University of Adelaide

October 30, 2015

Discipline of Statistics
School of Mathematical Sciences

THE UNIVERSITY
of ADELAIDE

Signed statement

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: ....

Contents

List of Figures

List of Tables

Acknowledgements

I must emphasise my gratitude to my supervisors, Professor Patty Solomon and Dr Chris Bagley. Patty, I have tremendous admiration for your statistical knowledge and thank you for your fantastic insights, guidance and wisdom along the way. Chris, your incredibly sharp eye combined with your patience when answering my questions is thoroughly appreciated.

Mum, Dad, Mel and Liana, I hope we can do away with the "are you done yet?" question. Thank you to you all for your encouragement from near and far, recently and formerly. I am under no illusions that without your combined support I would not be writing this now. I will be a better son, brother and partner now I promise. I extended this to friends who have been very understanding of my absenteeism. Thank you for your support as well.

Thank you to Chris Davies also for his initial work in his honours thesis that allowed me to hit the ground running. Many thanks to The University of Adelaide, specifically to the School of Mathematical Sciences and those I have had the most contact with at the Adelaide Proteomics Centre: Megan Penno, Vicki Clifton and Peter Hoffmann.

Since I have the floor, there are some more general sentiments I would like to make. I am grateful to exist in the time and location I do, and to be able to do what I love. My exclamation of "what a time to be alive!" is rarely sarcastic, albeit a poor attempt at humour. It would be remiss of me not to reference 'standing on the shoulders of giants' (but to complete the metaphor, in my case, rather than standing I might be sitting or even sliding off). I also wish to thank others that I have not met, those who get insufficient acknowledgement for what is a tremendous service to society: people that create and maintain publicly available software. Particularly the authors and contributors of R and TeX/LaTeX, software I have used extensively in this thesis.

Abstract

Proteomic spectra obtained from matrix-assisted laser desorption ionisation (MALDI) time-of-flight mass spectrometry (TOF-MS) are generated from the proteins and peptides present in serum obtained from blood. By ionising the proteins and resolving them in the mass spectrometer, data on the expression of proteins can be obtained, realised from the amplitude of signal for different mass to charge ratios. Of primary interest is the biological signal, in particular, the expression of proteins related to disease. In common with many ‘omic’ technologies, the raw spectra suffer from systematic errors due to technological artefacts and batch-effects, in addition to sample and biological variability. To negate these effects, novel application of genetic microarray pre-processing and analysis methods to proteomic TOF-MS data are presented. However, there are important differences between microarray and TOF-MS data which require consideration and non-trivial modifications to be successfully applied. One important difference between MALDI TOF-MS data and other high-throughput data, seldom addressed, is the high proportion of missing values.

The pre-processing of raw proteomic TOF-MS data needs to be undertaken prior to analysis and remains a mathematical and statistical challenge. Performed in distinct steps, pre-processing consists of signal smoothing, baseline correction, spectra normalisation, peak detection and peak alignment. An argument as to why the order of these steps is highly important is presented. Standard and novel data pre-processing methods are investigated and compared to optimise the process. Each step is given due consideration since the cumulative effects of substandard pre-processing can render subsequent statistical analysis highly unreliable.

Ultimately, the aim of proteomic MS is to analyse the protein profiles. Two different but related approaches to the analysis are undertaken. The first approach is to identify biological markers (biomarkers) that exhibit differential expression between disease groups. Identifying potential biomarkers for further research requires appropriate exploratory, visual and statistical modelling which is addressed in detail here. The second approach is to perform statistical discrimination between groups, a classical supervised learning problem. The ability of mathematical models to predict disease groups using differential biological signal provides insight into the plausibility of diagnostic tests. Methodologically, supervised learning is a multifaceted problem given that feature selection, model parameter optimisation, and the handling of the training and test data all contribute to the inference that can be made from the results. Empirical appraisal of the methods applied to the proteomic data are provided with the outcome of discrimination error as a quantitative benchmark.

A number of proteomic TOF-MS datasets with differing characteristics are used throughout this thesis to assess the validity of the methods presented. The detailed analysis of a murine model MALDI TOF-MS dataset has facilitated the discovery of potential biomarkers for gastric cancer. Correct classification of spectra to their respective disease group (gastric cancer or control mice) as high as 97.4% was achieved using supervised learning. The thorough treatment of all the differently behaved datasets contained in this thesis, starting from the raw data pre-processing steps through to the challenging process of identifying potential biomarkers, provides a comprehensive and best-practice pipeline to analyse real-world proteomic MS data.

Acronyms and abbreviations

For simplicity, many abbreviations will be used throughout this thesis. The abbreviation/acronym will appear in parentheses at the first occurrence of the phrase but the table below provides a comprehensive list for quick reference.

Abbreviation	Meaning
APC	Adelaide Proteomics Centre
C	The portable and compiled programming language
C8 beads	Alkyl group beads used in proteomic sample fractionation
CLSA	Continuous line segment algorithm
CLN	Cyclic LOESS normalisation
CRC	Colorectal cancer
CV	Coefficient of variation
(k)Da	(kilo)Daltons; ${}^{1}/_{12}{}^{\text{th}}$ of a carbon-12 atom's mass ($\sim 1.7 \times 10^{-27}$kg)
DNA	Deoxyribonucleic acid
DP	Dynamic programming
EQN	Empirical quantile normalisation
FDR	False discovery rate
FS	Fisher score
FWHM	Full-width at half-maximum
GC	Gastric cancer
GC-MS	Gas chromatography-mass spectrometry
GEE	Generalised estimating equation
GFCV	G-fold cross-validation, traditionally denoted k-fold
GLM	Generalized linear model
HM	Harmonic mean
IMAC-Cu	Immobilised metal affinity chromatography - copper
kNN	k-nearest neighbours
LC-MS	Liquid chromatography?mass spectrometry
LDA	Linear discriminant analysis
LME	Linear mixed effects
LOESS	Locally weighted scatterplot smoothing (local regression)
LSA	Line segment algorithm

Abbreviation	Meaning
MA	A transformation of paired minus vs. average log intensities
MAR	Missing at random
MALDI	Matrix-assisted laser desorption/ionisation
MCAR	Missing completely at random
MS	Mass spectrometry
m/z	Mass divided by charge: the x-axis of TOF-MS
μm	Micrometre (10^{-6} metres)
Nd:YAG	Neodymium-yttrium aluminium garnet (laser)
n_k	The number of patients/subjects in $k = 1, \ldots, K$ groups
nm	Nanometre (10^{-9} metres)
NW	Needleman and Wunsch (algorithm)
OOB	Out-of-bag
OLS	Ordinary linear least-squares (regression)
PC	Prostate cancer
PCA	Principal component analysis
PF	Pareto Front
PFDA	Pairwise fusion discriminant analysis
PLS	Penalised least squares (regression)
pH	Acidity/akalinity scale; hydrogen ion concentration metric
pmol/μL	Molecular concentration/microlitre; pmol $\approx 6 \times 10^{11}$ molecules
QDA	Quadratic discriminant analysis
R	The statistical programming environment
RDA	Regularised discriminant analysis
REML	Restricted maximum likelihood
RF	RandomForest
RNA	Ribonucleic acid
RUV	Remove unwanted variation
S2N	Signal to noise (ratio)
SAX	Strong anion exchange
SE	Structuring element
SELDI	Surface-enhanced laser desorption/ionisation
S-G	Savitzky-Golay
SnLp	Small-n Large-p (problem)
SVA	Surrogate variable analysis
SVD	Singular value decomposition
SVM	Support vector machine
SW	Smith and Waterman (algorithm)
TCN	TIC normalisation
TIC	Total ion current
TOF	Time-of-flight
$\mathrm{T_x}$	Treatment
UV	Ultraviolet
WCX	Weak cation exchange

Chapter 1

Disease, proteins and mass spectrometry

The motivation for this research is provided in this chapter. The required biological background and an outline of the technology used to generate proteomic mass spectra from sera is provided. Matrix-assisted and surface-enhanced laser desorption/ionisation time-of-flight mass spectrometry as a technology is one of many used in protein research; how this work is positioned in the broader context of protein research is additionally presented. The technological and mathematical challenges involved in the analysis of proteomic mass spectra are immense. Datasets used to highlight these challenges are outlined with brief reviews of their experimental designs. Having presented the aims, context and data, a basis for the pre-processing of the raw data and the data analysis will be established.

1.1 Motivation

Cancer is ultimately a disease of mutated genes in which unregulated proliferation of cells occurs without healthy occurrence of cell death. Some of the resulting cells then spread to other organs and inhibit normal function. Without treatment, the organism will die (Ruddon, 2007). However, cancer is a term covering different diseases with many biological pathways (Hanahan and Weinberg, 2011), symptoms and prognoses.

According to the World Health Organisation, 8.2 million people died in 2012 of cancers, accounting for 14.6% of all deaths worldwide (IARC, 2015b). The age-standardised cancer mortality rate in Australia is estimated at 80 (females) and 115 (males) per 100 000 people per year (IARC, 2015a). For colorectal cancer alone, the estimated disability-adjusted life-years per 100 000 people per year (the sum of years of life lost and years lived with a disability) is 337 for females and 258 for males in Australia and New Zealand (Soerjomataram et al., 2012). The direct health system cost of cancer in Australia is calculated at $4.5 billion per year (AIHW, 2013) alone, without the additional cost of illness taken into account (Rice et al., 1985).

Development of early stage, non-invasive and cost-effective screening tests for internal cancers, which are generally asymptomatic diseases, vastly improve a patient's prognosis and chances of survival (Schroder et al., 2009). The development of a non-invasive test using biological markers (biomarkers) in human sera with high sensitivity and specificity is the holy grail of biomarker discovery.

An example of an internal cancer in need of an accurate screening test is prostate cancer (PC). PC is a cancer of the male reproductive prostate gland, mostly affecting men 40 years and older. Currently there is no consensus on the use of the PC screening method using a prostate-specific antigen (PSA) threshold, from a blood test. PSA as a screening test is poor, having estimated sensitivity ranging from 20 to 40% and specificity ranging from 70 to 90% (Prensner et al., 2012). There is much debate about its effectiveness (Parpart et al., 2007) and even recommendations against its use (Moyer, 2012). There is high patient variability in baseline PSA levels, and thus the relationship of a PSA threshold to malignancy is not generalisable. A patient with high PSA levels may not have PC, and conversely, a patient with low PSA levels may have PC. A biopsy is performed if the screening test shows a high level of PSA; this in many cases is an uncomfortable, invasive test for PC (Issa et al., 2000).

With the emergence of mass spectrometry (MS) and other methods of biomarker discovery, it is hypothesised that considering the range of organic molecules in human sera, health care providers can better predict the cancer status of a patient and thus improve patient survival (Adam et al., 2002). Tumours in the body will shed proteins into blood or even reduce the abundance of proteins produced by healthy cells (Roy et al., 2011). These are proteomic biomarkers of interest. Biomarkers are not limited to proteins however, circulating nucleic acids, fragments of genetic code in the bloodstream, are examples of non-proteomic biomarkers showing promise in diagnostics as well (Swarup and Rajeswari, 2007).

The discrimination of proteomic mass spectra for prostate cancer (Petricoin et al., 2002b) and ovarian cancer (Petricoin et al., 2002a; Conrads et al., 2004) have been attempted previously. These studies have subsequently been discredited because of bias in the experimental design and errors in analysis (Sorace and Zhan, 2003; Baggerly et al., 2004, 2005; Solomon, 2009). There is still hope however that diagnostic and prognostic tests can be developed using time-of-flight (TOF)-MS on human sera but there are some hurdles to be overcome relating to the reproducibility and sensitivity of the results from the technology (Albrethsen, 2007; Gatlin et al., 2011).

Analysis of proteomic TOF-MS data is a challenging field for biologists and statisticians. The biological and technical aspects of MS induce systematic and other variation in the observed data, which must be identified and modelled directly (or removed) if underlying biological signals are to be detected. For this reason, MS studies must be well-designed at the outset. The mass spectrometry data can also be of high dimension, with hundreds of proteins measured for a single individual, resulting in the so-called *small-n, large-p* (SnLp) problem. The complexity of the statistical analysis of such data is compounded by the fact that the biomarkers of interest are usually low abundance signals and therefore difficult to distinguish statistically from systematic and random variation. Therefore a two-stage process is required to analyse the spectra. Firstly, the raw spectra must be ‘cleaned’ to remove background noise and other known (and unknown) sources of systematic variation. These initial steps are known as ‘data pre-processing’ (Chapters 2 and 3). Secondly, analysis of the pre-processed data is undertaken on the protein profiles generated using the corresponding disease state labels and other experimental factors (Chapters 4, 5 and 6).

Our approach to the analysis of the proteomic mass spectra will identify potential biomarkers. Two different approaches are taken: linear modelling (Chapter 4) and discriminant analysis (Chapters 5 and 6). Linear modelling uses the protein profiles as an outcome to regress on the experimental factors. Discriminant analysis, also referred to as classification or simply discrimination, conversely uses the protein profiles as the predictor and disease state as the outcome. The use of discriminant analysis has the following advantages in biomarker identification.

- Discrimination allows feature selection to identify differentially expressed proteins and potential biomarkers of interest.

- Discrimination provides an indication of the ability of the identified proteins to differentiate the groups, not singularly but as a profile of proteins.
- The correlation, dependence and interactions of the proteins can be uncovered in multivariate classification models.

There is no current standard pipeline to pre-process and analyse proteomic TOF-MS data. Proteomic mass spectrometry is a promising field that would be enhanced by improved experimental design and analysis. Such enhancements can contribute to the development of novel screening tests for cancer. This thesis hopes to contribute to statistical practice on proteomic mass spectra by comparing new and existing methods in order to make recommendations on 'best practice'.

1.2 Biological background

A human is a collection of many organs working in cooperation. The units that form organs are cells and are thus considered the fundamental unit of life (Campbell et al., 2006). The cell is surrounded by a membrane that controls the movement of material in and out. All cells contain deoxyribonucleic acid (DNA), which is the genetic information for life (Hartwell et al., 2008).

A cell is a system of many structures (cell membrane, organelles, ribosomes) and genetic information (DNA, ribonucleic acid, proteins). They are studied in cell biology and genetics respectively. Eukaryotic cells, the cells found in multi-cellular life additionally contain internal sub-membranes that partition the cell. The nucleus is a partitioned structure containing the genetic code.

There are important molecules that mediate in the function of the cells. Proteins are one of these. Proteins are complex polymers, a structure composed from chains of monomers called amino acids.

Cancer is caused by mutated genetics that result in unregulated proliferation of cells (from just one original mutated cell). Proteins are intimately involved in all genetic expression and the relationship will be discussed in the following sections. Hence, the relationship between proteins and cancer is apparent.

1.2.1 Amino acids and proteins

Protein comes from the word *proteious* which in Greek means *first place.* Proteins provide structure and are the functional units that act as enzymes to effect the chemical reactions in cells. A protein's function is governed by its three-dimensional structure, a result of its particular sequence of amino acids and their interactions with each other.

Amino acids are comprised of the elements carbon, hydrogen, oxygen, nitrogen and some also contain a small amount of sulphur. They all have a common core structure, as shown in Figure 1.1. All amino acids have a central carbon atom, called the alpha carbon. Attached to the left is the amino group (H - N - H) and attached to the right is the carboxyl (acid) group (O = C - OH). Thus, the name, amino acid. Below the alpha carbon in Figure 1.1, there is an 'element' R. The R is one of 20 possible molecular arrangements which determine the 20 different amino acids.

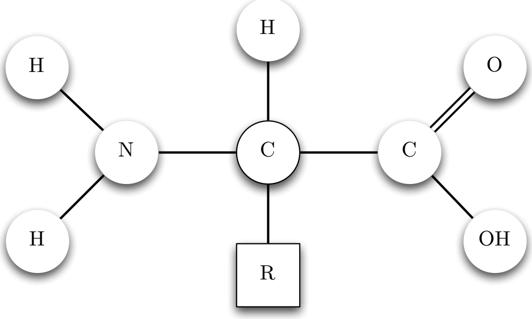


Figure 1.1: Elemental structure of an amino acid.

When strings of amino acids are chemically joined, they are called polymers. The chemical bonds that link the amino group of one amino acid to the carboxyl group of another amino acid are called peptide bonds. The result is a polypeptide. Most proteins are chains of a hundred or more amino acids that are a polypeptide or arrangement of polypeptides twisted into a unique three-dimensional shape that determines its utility. A polypeptide is not necessarily a protein. However, the terms peptide, polypeptide and protein are generally used interchangeably.

1.2.2 Proteins via genes

Proteins are coded by DNA. DNA consists of two strands, coiled together in what is called a double helix. The two strands are held together by nucleotides of which there are four types: adenine (A), thymine (T), cytosine (C) and guanine (G). These nucleotides are arranged in pairs between the two helices such that A is always paired with T and C is always paired with G.

DNA is the code for proteins and all other molecules making up an organism. For DNA to be expressed as proteins, the code must first be transcribed into ribonucleic acid (RNA) to then be translated into proteins.

The transcription stage is actioned by RNA polymerase, a protein that creates a single strand of nucleotides, an RNA strand, based on a sequence of DNA. Polymerase does this by assembling nucleotides using one side of the DNA sequence using the same pair-wise coding rules. The only exception is that RNA uses the nucleotide base uracil (U) instead of T.

As such, RNA is essentially a copy of one side of a segment of DNA. The location where RNA polymerase starts the transcription is at a specific sequence of nucleotides called promoter DNA. Similarly, RNA polymerase will end transcription of a DNA segment at a specific sequence of nucleotides called terminator DNA. Eukaryotic RNA is spliced to become messenger RNA (mRNA). The splicing removes the intron (non-coding) sequences to leave the exon (functional) sequences. The splicing may also select function specific exon sequences. In this way the same strand of RNA may become different mRNA sequences.

Translation of nucleotide triplets, codons, takes place to transition mRNA to protein. These codons map to the 20 amino acids (with the exception of three triplet permutations coding to a stop sequence). As there are $4^3 = 64$ permutations of triplets of the 4 nucleotides, many permutations code to the same amino acid. The mapping of codons to amino acids occurs with the use of the proteins: ribosomal RNA (rRNA) and transfer RNA (tRNA). The rRNA binds to the mRNA which in turn allows the tRNA to sequentially attach the amino acids according to the codon permutations. The tRNA is thus able to ‘grow’ a chain of amino acids (peptide).

The process of generating proteins from DNA is referred to as gene expression. This is the conversion of the genotype (the coded information in DNA) to the phenotype (the realised traits of an organism). The gene expression process, although a network of factors, is simplified as the central dogma of biology (Crick, 1970):

$$\text{DNA} \xrightarrow{\text{Transcription}} \text{RNA} \xrightarrow{\text{Translation}} \text{Protein.}$$

Thanks to the Human Genome Project (International Human Genome Sequencing Consortium, 2004; Pennisi, 2012), the human genome is currently estimated at 21,000 traditional (protein-coding) genes. The human proteome is some way off being fully characterised but is expected to contain hundreds of thousands to millions of post-translationally modified proteins (Anderson and Anderson, 2002; Jensen, 2004; Walsh, 2006). Genes will tend to code for proteins that are approximately 7kDa-1700kDa in mass but with post-translational modification, peptides can be fragments as small as a few amino acid masses with mass 500Da (Walsh, 2006; UniProt, 2013). Amino acids range from approximately 60Da-190Da.

1.2.3 Gene and protein expression

In eukaryotic cells there are many ways the expression of genes are controlled.

DNA packing is one such mechanism where DNA strands are coiled around proteins called histones. Transcription proteins are inhibited from attaching and transcribing the DNA because of the tight packing and folding of the DNA around the histone.

Transcription proteins such as RNA polymerase are moderated by activator and repressor proteins. Activator proteins attach to enhancer sequences, DNA segments that do not code for proteins themselves, to attract RNA polymerase and induce transcription. As a competing force, silencer DNA sequences attract repressor proteins that in turn inhibit the attachment of RNA polymerase that commences transcription of genes.

After a gene is transcribed, expression can be regulated with the breakdown of mRNA in the cytoplasm after it passes through from the nucleus of the cell. Even if the mRNA remains intact, translation to polypeptides does not guarantee a functional protein. Expression can be controlled if the polypeptide is not initiated or split into functional proteins.

Epigenetics is another class gene expression modification. The study of epigenetics is generally considered to be the study of non-DNA-based modification of gene expression (Berger et al., 2009). An example is DNA methylation where a methyl group (hydrocarbon) attaches to DNA nucleotides (specifically, A and C nucleotides) and in turn changes the regulation and expression of the corresponding sequence.

1.2.4 Protein function, sera and cancer

Biomolecules are created in chemical processes with the assistance of enzymes. These chemical processes are effectively shaped by genes which encode what is to be made (Hartwell et al., 2008). The sequence of these chemical processes are directed by the environment and the activities of other biomolecules made within the cell or by other cells. These events and biomolecules are the study of biochemistry, physiology, genetics and cell biology.

With these inquiries into biological processes, it is important to ask what the composition of the elements are, what their purpose is and how this purpose is achieved. Understanding how these molecules work and the chemical processes involved in disease leads to the process of biomarker discovery.

Biomarkers are biological molecules that are indicative of that structure's state or condition. Biomarkers can range from a metabolite to a network of genes. Biomarker discovery is not only the process of finding such molecules but the identification and characterisation of these molecules to understand the pathways of their related interactions. In a proteomics context, the purpose of biomarker discovery is to find whole or naturally occurring subsets of proteins that are related to a disease state. The ultimate purpose being the discovery of biomarkers that exist on a biological gradient with the underlying spectrum of severity or prognosis of disease.

The relationship of different proteins to a disease varies greatly. Enzymes are predominately proteins themselves, thus an enzyme might be indicative of a cancerous growth's chemical reactions. Many hormones are proteins as well; hormones function as inter-cell communicators and regulators. Protein hormones, as opposed to steroid and other hormones, generally move between cells via the membrane of the cell and can find themselves in the circulatory system. In this way, blood will contain biomarkers of cancer.

Biomarkers may not be directly related to the disease of interest but may be precursors, derivatives or by-products of the disease's biological pathway. Additionally, these proteomic biomarkers may not be functional proteins but fragmented proteins degraded by proteases. As such there will be small signals of biologically relevant biomarkers amongst the large noise of other proteins.

As this thesis is concerned with the proteomic biomarkers in serum, the constituents of blood are considered in Figure 1.2. The separation of plasma from the other components of blood allows serum to be studied. Within the serum component of Figure 1.2, proteomic biomarker candidates are listed.

1.2.5 Proteomics

Proteomics is one of many 'omics' disciplines that have arisen from modern technologies to study 'omes' on large scales, many of which can be found in the field of bioinformatics. *Omics* and *omes* are modern suffixes to biological sub-specialties (Hotz, 2012). 'Omics' is the study of a particular part of biology where 'ome' is the corresponding set of objects in the field (Lederberg and McCray, 2001). Examples are genomics for genomes and metabolomics for metabolomes. Sub-domains of these areas are also defined for research, such as the human tumour-related plasma proteome (Omenn et al., 2005; Hortin, 2006; Villanueva et al., 2006; Vizcaino et al., 2013). The aim in this thesis is to examine the human and animal proteome to find indicators of disease.

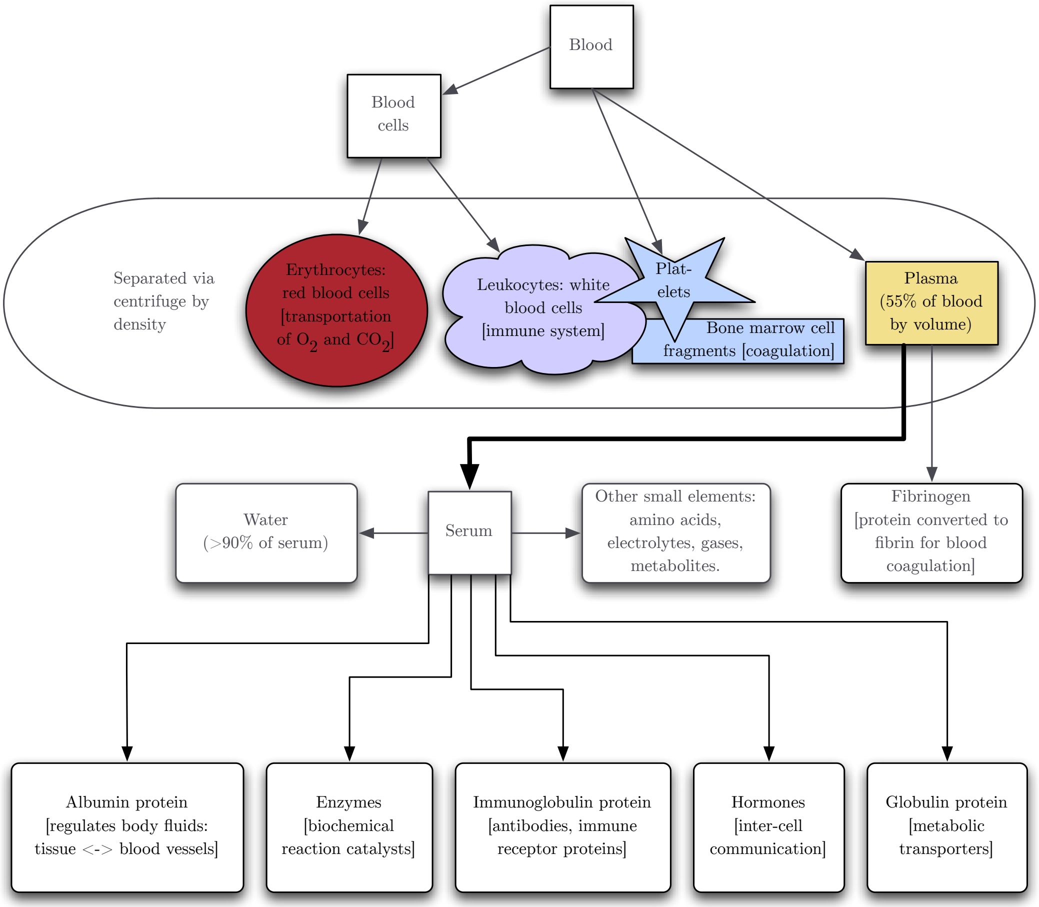

Blood
Blood cells
Separated via centrifuge by density
Erythrocytes: red blood cells [transportation of O_2 and CO_2]
Leukocytes: white blood cells [immune system]
Plat- elets
Bone marrow cell fragments [coagulation]
Plasma (55% of blood by volume)
Water (>90% of serum)
Serum
Other small elements: amino acids, electrolytes, gases, metabolites.
Fibrinogen [protein converted to fibrin for blood coagulation]
Albumin protein [regulates body fluids: tissue <-> blood vessels]
Enzymes [biochemical reaction catalysts]
Immunoglobulin protein [antibodies, immune receptor proteins]
Hormones [inter-cell communication]
Globulin protein [metabolic transporters]

Figure 1.2: The components of blood, compiled using a variety of sources based on information found in Litwack (2008); AABB (2013); Dominiczak and Fraser (2014).

The new age of high-throughput technology requires quantitative expertise and collaboration between many fields. A feedback loop exists where biological knowledge and hypotheses influence the mathematical models and vice versa, where the high-throughput approach to data collection requires the statistician to inform the biologist of the relationships between variables and other insights based on the data. Such a relationship is outlined in Figure 1.3.

As one might expect, such endeavours are not straight-forward. Because of the complexity of the protein interactions in animals, a single technology will not always identify the entire network of relationships between proteins, metabolites, enzymes and so on. Listed below are just some of the considerations from a bioinformatics perspective.

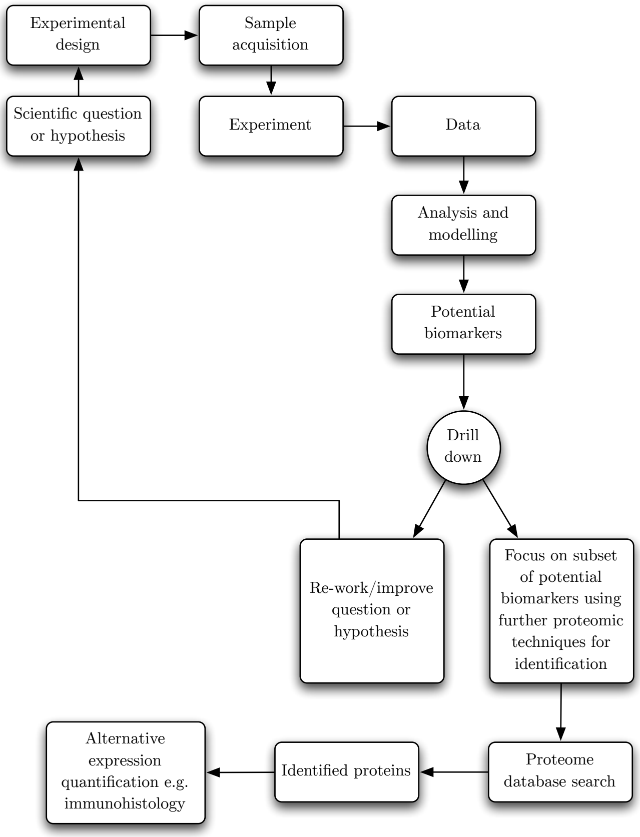

Experimental design
Sample acquisition
Scientific question or hypothesis
Experiment
Data
Analysis and modelling
Potential biomarkers
Drill down
Re-work/improve question or hypothesis
Focus on subset of potential biomarkers using further proteomic techniques for identification
Alternative expression quantification e.g. immunohistology
Identified proteins
Proteome database search

Figure 1.3: The context of biomarker discovery, as envisaged in this thesis.

- Peptides may be up- or down-regulated due to a disease and this is one of many reasons for differential expression of proteins. There may be temporal, spatial or other confounding factors influencing the up- or down-regulation of the peptides. For example, heat may induce the expression of heat shock proteins or the experimental design may induce batch effects skewing the results.
- Proteomic profiles contain peptides that are post-translationally modified and the potential biomarkers may be variants of differently transcribed versions of the same gene (isoforms) or protein fragments derived from the original protein.
- If a peptide is differentially expressed, this may be a result of a differentially expressed original protein or a result of the biochemistry surrounding the disease that alters the distribution of fragmented peptides of the original protein.
- Potential biomarkers may follow a biological gradient with respect to the disease, where an increase of the severity or progression of the disease results in higher or lower expression of the biomarker(s).

1.2.6 Biomarker discovery

There are two broad means of biomarker discovery: display and identification. The former allows many biomarkers to be presented at the same time while the latter can be considered a drill-down of the former where individual (or a sub-group of) proteins are examined (Figure 1.3).

Two-dimensional (2D) electrophoresis is an example of a display method of biomarker discovery, developed in the 1970s. This technique for studying biomarkers separates polypeptides generally greater than 10kDa on two characteristics, and requires two steps. The first step separates polypeptides in a gel by their isoelectrical properties (more specifically, the pH at which the polypeptide has no net electrical charge). The second step involves a further separation of the polypeptides in the perpendicular direction based on the polypeptide's weight. This is performed by binding anionic detergent to the proteins and the rate at which the polypeptide can move through the electric field is inversely related to their mass. The resulting 2D image is used to compare the distributions of biomarkers between diseased and non-diseased individuals (Gharbi et al., 2002).

Another commonly used technique for display is mass spectrometry. This method creates a spectrum of mass divided by charge (m/z) values plotted against the corresponding amount of polypeptide (intensity), typically for polypeptides less than 20kDa. An example of a mass spectrum is shown in Figure 1.4.

MS can be considered a combination of two parts: a mass analyser and an ionisation process. There are four basic types of mass analysers (Glish and Vachet, 2003).

(1) Linear time-of-flight (TOF) - uses acceleration of charged biomolecules to deduce the m/z values by the time they take to reach the detector.

(2) Reflectron TOF - which is similar to linear TOF with an additional 'ion mirror' (electric field) that reflects the charged particles in roughly the opposite direction after the initial TOF section.

(3) Quadrupole - uses four rods that are given specific radio frequencies and voltages to only allow polypeptides with conforming m/z values through.

(4) Quadrupole ion-trap - as the name indicates, instead of passing the ionised molecules through the instrument, the ions are trapped in an array of electrodes. Some of these electrodes are alternated to generate either complex 2D or 3D movement of the molecules via changing electric fields before being ejected to resolve the ionised molecule's mass.

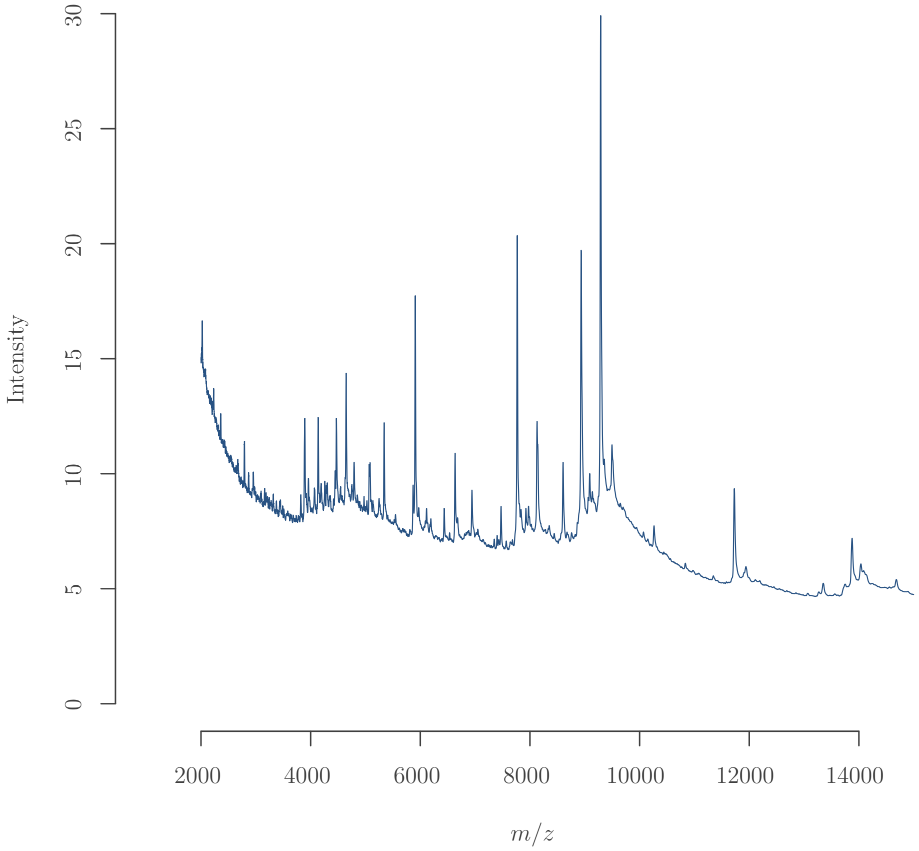


Figure 1.4: An example of a mass spectrum (m/z vs intensity): a serum derived SELDI TOF-MS raw spectrum from the Adam et al. (2002) study.

There are three types of ionisation processes that will be examined in more detail in §1.2.8.

(1) Matrix-assisted laser desorption/ionisation (MALDI).

(2) Surface-enhanced laser desorption/ionisation (SELDI).

(3) Electrospray ionisation.

Linear TOF-MS (using MALDI or SELDI ionisation) is a widely used and effective method for biomarker discovery and will be the focus of this thesis. A more detailed explanation of TOF mass analysers is given in §1.2.7.

An example of biomarker discovery for protein characterisation is tandem mass spectrometry (also referred to as MS/MS). MS/MS uses a sample with an isolated polypeptide. This sample is subject to proteolysis, normally via a protease to cleave the polypeptides. The constituent parts of the original polypeptides are put through a mass spectrometer to produce a spectrum of the peptide fragments. From this spectrum, the amino acid sequence of the original polypeptide can be deduced to identify the parent polypeptide (Rudnick et al., 2010). Another example of biomarker identification is high performance liquid chromatography that separates polypeptides for characterisation. This process involves passing a solution containing the analyte under high pressure through tubes with a solid matrix inside that separates peptides by adsorption characteristics (Shi et al., 2004).

1.2.7 Linear TOF-MS

This thesis is concerned with the use of linear TOF-MS to detect proteomic biomarkers of internal cancers and other diseases. Using the TOF of molecules to deduce their mass was theorised in 1946 and first used for protein research in the 1950s and 1960s (Yates III, 2011; Borman et al., 2003; Wiley and McLaren, 1955).

The basic elements of the TOF design are shown in Figure 1.5. Such an instrument acts on a charged particle (e.g., a protein with a positive or negative charge). Here, a positively charged particle is illustrated at the positively charged presenting plate on the left side of Figure 1.5. The particle starts in an electric field and accelerates from the source or presenting plate. Once the charged particle reaches the end of the electric field, it has reached a velocity it will travel through the *free-flight tube* (the section with no electric field in Figure 1.5) in which there is a vacuum.

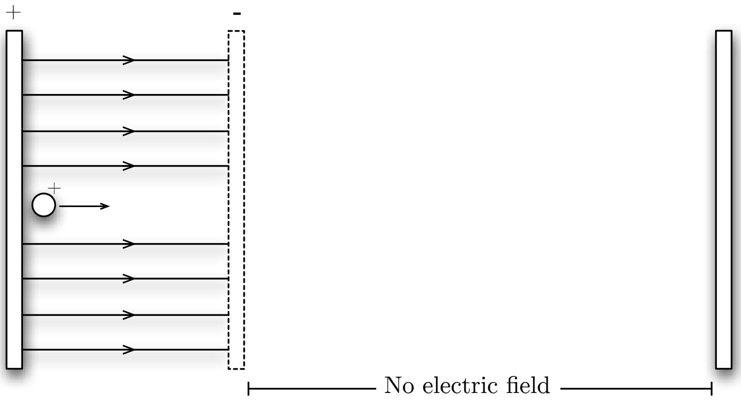


Figure 1.5: An illustration of the basic construct of TOF-MS.

The reason this is called TOF-MS is because the time-of-flight, t, of a particle is related to its mass, m. The masses of proteins are of interest to the researcher.

The particle depicted in Figure 1.5 represents one of many proteins from the serum sample that travel down the TOF-MS to create a mass spectrum. However, proteins are charge neutral so they need to be given net electric charge. How this is performed, while keeping their mass and structure intact, is outlined in §1.2.8.

Relationship between mass, charge and time-of-flight

The time-of-flight of a charged particle is proportional to the square root of the mass divided by the charge, z, of the particle, $t \propto \sqrt{m/z}$. An outline of how this relationship is deduced is available in Merchant and Weinberger (2000).

As inference is made on the mass of the proteins indirectly and there may be peptides starting with non-zero velocities, calibration of the TOF system is required. Generation of a time (x-)axis and an intensity (y-)axis requires solving $\sqrt{m/z} = c_0 + c_1 t$ for the constants c_0, c_1 i.e. solve $m/z = (c_0 + c_1 t)^2$.

In theory, the relationship between t and $\sqrt{m/z}$ can be derived using the values of the system (electric fields, distance), but calibration for c_0, c_1 in the relationship $\sqrt{m/z} = c_0 + c_1 t$, using at least two peptides of known m, z values is the preferred method (Merchant and Weinberger, 2000). The peptides chosen in the calibration should be in the range of m/z values to be analysed so that no extrapolation is required in the resulting mass spectra.

1.2.8 Ionisation of proteins

For linear TOF-MS to accelerate proteins in the system, the proteins must be ionised (i.e., given a net electric charge). The creation of ionised proteins without destroying the structure and mass of the proteins has been a trial and error process (Beavis et al., 1989a,b,c; Beavis and Chait, 1990).

One of these ionisation processes used to generate the data analysed here, MALDI, requires the sample of interest to be embedded in a *matrix* (a crystallised acid) so that the contained peptides non-covalently bond with the matrix (i.e., do not share electrons with the matrix molecules). The matrix is chosen because of its ability to absorb electromagnetic radiation while protecting the proteins from being damaged. The matrix solutions α-cyano-4-hydroxycinnamic acid, 2,5-dihydroxybenzoic acid (gentisic acid) and 3,5-dimethoxy-4-hydroxycinnamic acid (sinapinic acid) have been used (Beavis and Chait, 1996). By firing short pulses of laser light (i.e., ultraviolet radiation, UV, in such circumstances), the matrix becomes volatile. This ejects the analyte's molecules in the gas phase, allowing the peptide to be desorbed (ejected, and thus possibly starting with a velocity > 0) with one or, less likely, two protons (hydrogen ions, H^+) attached via transfer with the matrix. This is also referred to as soft ionisation as the proteins are generally kept intact without fragmentation.

Fractionation is used to separate biomolecules within the analyte that are suitable for analysis, prior to being embedded in the matrix. Different fractionation methods will attract different subsets of biomolecules of interest. In this way, the choice of fractionation technology is an area of experimentation in itself. Magnetic beads that bind with peptides (whilst removing components counter-productive to MS) are facilitated using various properties. The types of magnetic beads used in the experimental data analysed in this thesis are hydrophobic interaction (C8), immobilised metal affinity chromatography (IMAC) and weak cation exchange (WCX) beads. Only a small amount of the (post-fractionated) analyte is required for the analysis, of the order of pmol/μL in a purified water-diluted matrix mixture. The solution is dried and crystallised on the target plate in the mass spectrometer.

The lasers used on the matrix solution are in the UV range (wavelength $\lambda < 400$nm). Nitrogen lasers with wavelength $\lambda = 337$nm have been used or neodymium-yttrium aluminium garnet (Nd:YAG) lasers that produce larger wavelengths of (infrared, wavelength $\lambda = 1064$nm) electromagnetic radiation can be used in conjunction with frequency tripling ($\lambda = 354$nm) or quadrupling ($\lambda = 266$nm). The wavelength of the laser light is dependent on the matrix used. Both types of lasers are used in pulses; the Nd:YAG laser uses a Q-switching method. By pulsing the lasers at the matrix, the molecules are briefly excited, producing ionised proteins that are accelerated towards the detector as discussed previously. By firing short and soft electromagnetic radiation at the matrix, intact proteins are retained. As the laser is pulsed systematically, the recordings by the detector for each pulse can be summed together to create a mass spectrum. The duration of the pulsed laser light is of the order of nanoseconds (Beavis and Chait, 1996).

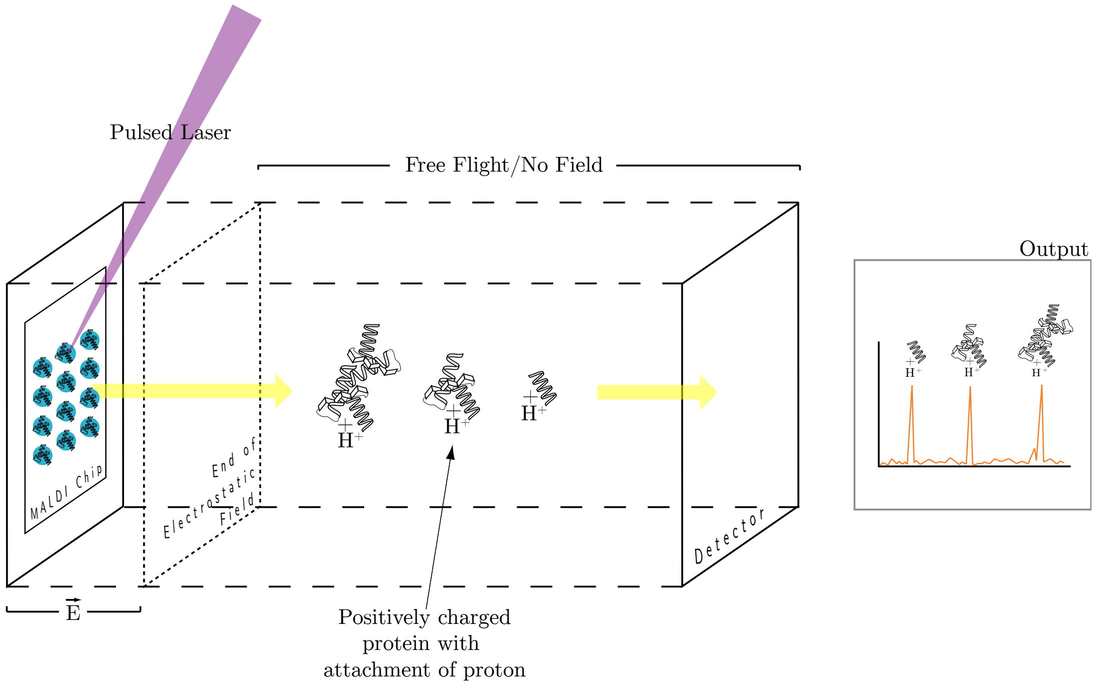

Pulsed Laser
Free Flight/No Field
Output
MALDI Chip
End of Electrostatic Field
H+
H+
H+
H+
H+
H+
Detector
$\vec{E}$
Positively charged protein with attachment of proton

Figure 1.6: A schematic of a MALDI TOF-MS system.

This method of ionising polymers is termed MALDI and has been used successfully on synthetic peptides and a range of biological molecules (Merchant and Weinberger, 2000). For clarity, a representation of the MALDI TOF-MS system from ionisation to the generated mass spectrum is shown in Figure 1.6.

The development of an alternative but similar method, SELDI, followed the article of Hutchens and Yip (1993). The SELDI technology was commercialised by Ciphergen Biosystems in 1997. The Ciphergen SELDI ProteinChip Array System technology has since been taken over by Bio-Rad Laboratories (Bio-Rad Laboratories, 2010).

SELDI works without a matrix solution. The sample is placed on the SELDI chip that captures the peptides using one of the various chemical affinities available, similar to those used by the C8 beads in the MALDI process. The sample is left on the chip for an ‘incubation’ period before the remaining sample is washed off. Similarly to MALDI, the chip is irradiated with a laser and the surface absorbs the radiation and desorbs the attached proteins. There is a range of SELDI chip-types to affinity capture proteins, including hydrophobic, hydrophilic, anionic, cationic, metal ion and combinations therein (Issaq et al., 2002). Experimentation to find the best chip to produce a signal is required for different uses.

SELDI has been used less frequently in recent times because of its lower spectrum resolution (Gemoll et al., 2010; Albrethsen, 2011) as well as its lack of reproducibility (Semmes et al., 2005). The mass accuracy (peak drift) of SELDI TOF-MS is also inferior. Mantini et al. (2010) outlines a peak drift of 2000 parts per million (0.2% of the peptide mass) for SELDI TOF-MS and 300 parts per million (0.03% of the peptide mass) for MALDI TOF-MS.

Currently, MALDI and SELDI TOF-MS are limited to low molecular weight proteins and protein fragments. In most cases these are peptides less than 20kDa in the proteome (Karbassi et al., 2009; Terracciano et al., 2009). It is also worth noting that peptides less than 1-2kDa are generally hard to resolve because of the noise generated by rogue-charged matrix particles; this low-weight component of spectra is generally removed prior to analysis (Glish and Vachet, 2003).

The electrospray technique is an alternate ionisation process that passes a solution with the analyte through a very small tube (with diameter of the order of 100μm) with an electric potential difference across it. By electrostatically spraying the solution, evaporation of the solution occurs, leaving individual molecules of analyte to enter the MS. The analytes will disassociate with the solution, resulting in a spectrum of the molecules resolved according to their mass to charge ratio. Electrospray ionisation differs from MALDI/SELDI techniques in that MALDI/SELDI will generally produce singly-charged ions, whereas electrospray will produce a range of charges resulting in separate peaks of m/z values of the same molecule. Electrospray ionisation suffers from some limitations: more of the analyte is required than for MALDI/SELDI techniques (Glish and Vachet, 2003) and is also sensitive to salt content in biological samples. However, the latter can be overcome with 'in-line' technologies (Chen et al., 2011).

1.3 MS Data

A single TOF-MS spectrum is measured as an array of positive intensity values for discretely measured m/z-values, although the underlying profile can be considered continuous. This is why spectra are depicted as lines, as shown in Figure 1.4 (for example).

An observed spectrum i, from $i = 1, 2, \dots, n$, can be considered as a vector $\boldsymbol{f}_i = (f_i(1), f_i(2), \dots, f_i(T))$ where each $f_i(t) \in \mathbb{R}^+$ is an expression value for the $t = 1, 2, \dots, T$ discrete m/z-values. The m/z values are not equidistant because of the non-linear relationship between TOF and m/z; m/z values will be closer together for low m/z-values and further apart at larger m/z values.

It is not uncommon for subsets or the entire set of m/z-values to be different from one spectrum to another, even for the same range of m/z. It is desirable to have mass spectra with common m/z-values and this is trivially achieved using interpolation of the $\boldsymbol{f}_i$ values for some common set of m/z values (Li et al., 2011; Gong et al., 2012), $t = 1, 2, \dots, T$.

The observed vectors, $\boldsymbol{f}_i$, require pre-processing, a non-trivial task outlined in Chapters 2 and 3. The aim of the pre-processing is to refine and condense the spectra into a peak expression matrix, E. The data E consist of P peaks that are common to the n spectra. For each spectrum, the value for a given peak, if detected in that particular spectrum, is a representative value of the protein expression. This can be peak area or peak height for example. This will be discussed further in Chapter 3. Discussion of the treatment of missing values when peaks are not found is addressed in Chapter 4.

The peak expression data take the form,

$$E = \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1P} \\ e_{21} & e_{22} & \dots & e_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ e_{n1} & e_{n2} & \dots & e_{nP} \end{bmatrix},$$

where $e_{ip} \in \{\mathbb{R}^+, \texttt{missing}\}$ is the expression of peak $p = 1, 2, \dots, P$ for spectrum $i = 1, 2, \dots, n$. It is standard to $\log_2$-transform the non-missing values to provide roughly symmetric peak expression distributions (Morris et al., 2005) and has been shown to be variance-stabilising (Wolski et al., 2005).

The peak expression data have been labelled here as E because they can be considered as either a matrix of outcomes, $E = Y$, or a matrix of predictors, $E = X$, depending on the context.

Chapter 4 considers the peak expression matrix as an outcome variable of a linear model, Y, with experimental variables such as the K disease groups forming a predictive design matrix, X. On the other hand, Chapters 5 and 6 consider the peak expression matrix as predictive observations, X. The predictive peak expressions seek to estimate group membership of spectra to the K disease classes as an outcome, via supervised learning techniques.

1.3.1 Synthetic data

Prior to the introduction of the experimental data used in this thesis, it is useful to explain the form of the noise and non-biological signal one needs to remove from the data before effective analysis of the spectra can take place; the practice of pre-processing. A simple model of a mass spectrum i takes the form,

$$f_i\left(t\right) = B_i\left(t\right) + N_i \times S_i\left(t\right) + \epsilon_i\left(t\right), \tag{1.1}$$

where t is the time-point/mass, f_i is the $\log_2$ transformation of the realised signal, B_i is the baseline signal, N_i is the normalisation constant, S_i is the true signal and ϵ_i is the additional noise (Morris et al., 2005).

Data based on the model in Equation (1.1) were randomly generated by Morris et al. (2005) and are available publicly.[1] These data are useful for assessing pre-processing techniques as the population parameters are known. For each virtual experiment of 100 spectra, a set of 150 virtual proteins were generated. The log of protein masses were randomly generated from a common normal distribution. Within each spectrum, whether a protein peak appears and the corresponding peak expression were randomly generated from Bernoulli and normal distributions, respectively. The parameter values of these distributions are random variables themselves, where the population parameters were derived from previous proteomic MALDI TOF-MS experiments. The generated values (please see Morris et al. (2005) for full details) were then submitted to a virtual MALDI-TOF instrument developed by Coombes et al. (2005) that generates the spectra. An example spectrum from these data is shown in Figure 1.7.

[1] Available at `http://bioinformatics.mdanderson.org/Supplements/Datasets/Simulations/index.html`

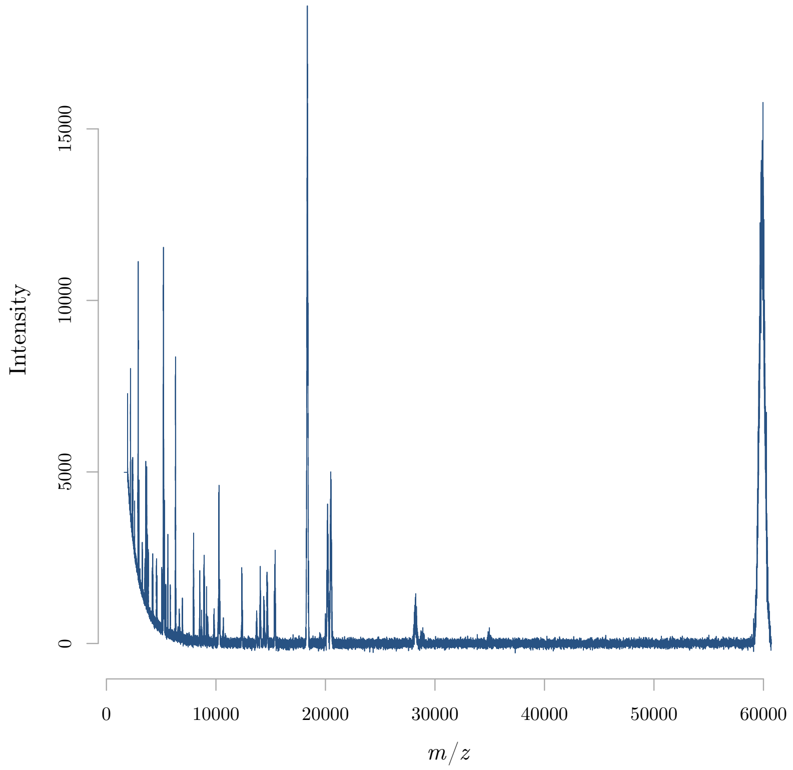


Figure 1.7: An example spectrum from the Morris et al. (2005) generated data.

1.3.2 Proteomic MS for cancer classification

The first study to claim successful use of proteomic mass spectrometry for differentiating cancer from non-cancer patients was Petricoin et al. (2002a). It reported 100% sensitivity and 95% specificity in detecting ovarian cancer in women. A follow-up study on ovarian data produced by the same group was also published (Conrads et al., 2004). A study by the same group, this time on prostate cancer, was also published (Petricoin et al., 2002b). The group's work received extensive exposure and prompted a push to generate a commercial proteomic ovarian cancer screening test. However, papers showing experimental design bias and analysis errors in these studies were later published (Sorace and Zhan, 2003; Baggerly et al., 2004, 2005; Solomon, 2009). It is generally accepted the Petricoin group studies are flawed (Alexandrov et al., 2009) and the actual diagnostic ability of classifiers used in these studies are no better than classification by chance (Baggerly et al., 2005).

This thesis investigates the data outlined in the coming sections with the following aims.

(1) To find potential biomarkers.

(2) Assess the state of current technology in distinguishing between disease groups. This refers to both TOF-MS technology and mathematical methodology.

(3) Validate potential methods to pre-process TOF-MS data.

1.3.3 Adam et al. (2002)

The paper published by Adam et al. (2002) was an apparently promising study which claimed to have excellent sensitivity and specificity as a potential diagnostic tool for prostate cancer. The data were generated using the SELDI technology popular at the time.

Blood samples were collected from the Virginia Prostate Center Tissue and Body Fluid Bank. Only samples prior to treatment for PC patient were used to avoid protein signal as a result of treatment not disease state. The study consisted of 97 age-matched healthy male controls (Cont), 92 benign prostate hyperplasia males (BHyp), 99 organ confined PC males (CanA) and 98 non-organ confined PC males (CanB). Some demographic characteristics of the patients described in the paper are summarised in Table 1.1. How age-matched controls were selected is not established in the paper. The mean and maximum ages for the control group were much lower than for the other three groups (Table 1.1). An imbalance of race between the groups is evident from Table 1.1. Any differential expression between the disease groups of mass spectrum profiles could be influenced by these factors.

Table 1.1: Summary of subjects in the Adam et al. (2002) study.

		Age			Race		
Class	n	Min	Mean	Max	Caucasian	African-American	Other or Unknown
Cont	97	51	60	70	50%	50%	0
BHyp	92	48	67	86	36%	2%	62%
CanA	99	50	71	89	77%	20%	3%
CanB	98	44	69	87	82%	16%	2%

The ProteinChip Array System of Ciphergen Biosystems was used to generate the spectra. The IMAC-Cu chip was deemed the most successful chip for affinity capture of the proteins after testing chip chemistry suitability. The SELDI TOF-MS spectra were limited to 2000-40000 on the m/z-axis for their analysis.

In the analysis of Adam et al. (2002), the CanA and CanB groups were combined into one PC group. The analysis focussed on the differentiation in the PC, BHyp and Cont groups. This is consistent with the aim of finding biomarkers to predict the presence of PC. However, it is an important exercise to differentiate the CanA and

CanB groups to determine the robustness of the methods used and the potential of protein identification and characterisation of biomarkers with differential expression between the two groups and types of cancers.

As the subjects were of known disease status, the aim was to create a model to predict disease status via a supervised classification method, using a training and test dataset drawn from the 386 subjects. The data available for analysis in this thesis were the 326 training samples resulting with sample sizes for the Cont, BHyp, CanA and CanB groups of $n_{\text{Cont}} = 81$, $n_{\text{BHyp}} = 78$, $n_{\text{CanA}} = 84$ and $n_{\text{CanB}} = 83$, respectively.

The pre-processing of the spectra was undertaken using software produced by Ciphergen Biosystems and an average of 81 peaks per spectrum were found (no variability is provided). Using these peaks, a classification tree was created using the training data. For the test data used in the Adam et al. (2002) study, the classification tree correctly classified 15/15 (100%) of the Cont, 14/15 (93%) of the BHyp, 12/15 (80%) of the CanA and 13/15 (87%) of the CanB patients.

To demonstrate the reproducibility of the SELDI TOF-MS process, spectra were generated for randomly selected duplicate samples at a later date. There is some confusion in the paper whether this was months, up to a year later or 18 months later. The number of duplicate samples that were selected was not outlined. These duplicates were then processed and placed through the classifier. The authors claim all spectra were assigned to the ‘appropriate’ node. It is ambiguous whether ‘appropriate’ node means correct classification or classification to the same node as the original replicate.

The positive results from the Adam et al. (2002) study promoted further investigation. Grizzle et al. (2003) outlined a three-stage comprehensive study into the use of SELDI TOF-MS for prostate cancer classification.

(1) Test platform reproducibility in a multi-institutional setting with three subparts.

    (a) Standardise protocols.
    (b) Inter-institutional reproducibility tested with coefficient of variation and intra-class correlation of peak’s intensity and location.
    (c) Reproducibility of spectra in reference to the central site (Eastern Virginia Medical School).

(2) Reproducibility of classification using spectra from multiple sites.

(3) Validation of early detection algorithms on spectra using a ‘comprehensive’ protocol using results of stages 1 and 2.

The first stage and assessment of reproducibility were satisfactory (Semmes et al., 2005) and progressed to the second stage. Two papers were published for the second stage (McLerran et al., 2008a,b). The first paper identified some bias relating to sample storage affecting results and the second paper states in the title "SELDI-TOF MS whole serum proteomic profiling with IMAC surface does not reliably detect prostate cancer". From this the authors did not pursue the third stage as requirements of the second stage were not met. Despite this, the Malik et al. (2005) and Drake et al. (2006) papers that stemmed from the earlier SELDI TOF-MS prostate cancer detection studies (Adam et al., 2002; Qu et al., 2002), identified the 8.9kDa peak that was over-expressed in PCA cases as apolipoprotein A-11 (ApoA-11). O. J. Semmes, G. Malik and M. D. Ward applied for a patent of this biomarker in 2006 and were awarded the patent in 2010.[2]

1.3.4 de Noo et al. (2006)

The de Noo et al. (2006) study was performed to test the reliability of TOF-MS as a diagnostic tool for disease using human sera. In this case colorectal cancer (CRC) was the disease under investigation. It was set up to address concerns discussed in §1.3.2 regarding batch effects, specifically 'day-to-day' and 'chip-to-chip' effects (de Noo et al., 2006).

The CRC samples, of roughly equal numbers of male and female patients, were taken one day prior to surgery for their condition. The 66 samples only resulted in 63 spectra because of inadequate profiles observed by manual inspection. Using histologies that confirmed the malignancy, malignant tumour classification (TNM) staging was assessed. Most colorectal patients were TNM stage 2, where the least severe is stage 1 and the most severe is stage 4 (lymph node metastasis).

There were 50 control patients resulting in 50 spectra. No information is supplied on whether age- and sex-matching occurred. The mean age and sex ratios across groups suggest it may not have (Table 1.2). Blood samples of cancer patients were collected over 27 months (October 2002 to December 2004) while control blood samples were collected over three months (October 2004 to December 2004). Ideally, the control blood samples would be taken over the same time period to negate issues of confounding.

Peptide fractionisation was performed using hydrophobic C8 magnetic beads. Three MALDI plates (Bruker Daltonics) were used to spot the fractionised samples. Each sample was spotted in quadruplicate. From the text and tables in the paper it can be deduced that replicates were limited to the same chip, while patients were randomised to chips resulting in roughly equal numbers of control patients and TNM

[2] Application No: 11/794838; Patent No: US7811772B2, 12 October 2010.

Table 1.2: Demographic information for the de Noo et al. (2006) study.

	Mean age (min-max)	Male	Female	Total
Canc	62.2 (32.6-90.3)	31	32	63
Cont	49.7 (25.9-76.6)	21	29	50

stage patients across the chips. Each chip was prepared and run to extract spectra on consecutive days.

The same process was repeated (sample preparation and spectra generation) a week later. Unfortunately these data are not available to us. The data available are the averaged quadruplicate spectra for each patient and the spectra are pre-processed, 113 spectra in total.[3]

The analysis of the spectra was performed using linear discriminant analysis with nested leave-one-out cross-validation resulting in detection of CRC with 95.2% sensitivity and 90.0% specificity as seen in Table 1.3. Alexandrov et al. (2009) revisited the data and were able to classify the spectra with 98.4% sensitivity and 95.8% specificity using wavelet pre-processing methods and support vector machines in nested five-fold cross-validation.

Table 1.3: Classification results for the de Noo et al. (2006) study.

	Estimated class		Classification
True class	Canc	Cont	error
Canc	60	3	
Cont	5	45	$^8/_{113} = 0.071$

1.3.5 Asthma studies

The Adelaide Proteomics Centre (APC) provided two asthma datasets for analysis that will be referred to as asthma1 and asthma2. While not cancer data, these MALDI TOF-MS data allow pre-processing validation and analysis for two experimental groups, as well as the opportunity to work closely with the experimenters.

It is estimated that 12% of mothers in Australia have asthma (Kurinczuk et al., 1999) and this has been linked to reduced birth weight in children (Murphy et al., 2002). These datasets are motivated by the work of Murphy et al. (2005, 2006). It

[3] The data from the de Noo et al. (2006) study used in this thesis were obtained from `http://www.math.uni-bremen.de/~theodore/MALDIDWT`, made available as the supplementary material for Alexandrov et al. (2009).

is hypothesised that the regulation of immune and vascular cells differ in male and female foetal development (Enninga et al., 2015) and therefore the involved proteins are likely to be differentially expressed. The aim of these asthma datasets was to find biologically relevant differences in the proteomic profiles of serum from pregnant mothers in asthma, asthma treatment and foetal sex groups.

Maternal plasma was sampled from $n = 30$ pregnant mothers at 30 weeks gestation who had single births. There were 20 mothers with asthma and 10 non-asthmatic mothers. Within the asthmatic mothers, half used glucocorticoid inhaled steroids. Sex of the child was evenly balanced within each of the sub-groups. Unfortunately the age of the mothers and other demographic information is not available. Recruitment was performed at the John Hunter Hospital antenatal clinic and ethics approval was obtained from the Hunter Area Health Service and University of Newcastle Human Research Ethics Committees.

The researchers were interested in proteins differentially expressed between the following groups of mothers, in order of priority.

(1) Sex of the child.

(2) Sex of the child and asthma status of the mother subgroups.

(3) Sex of the child, asthma status of the mother and steroid use subgroups.

This thesis will only address the primary aim above; more comment will follow in the remaining sections.

MS data were generated by the APC in-house using their MALDI TOF-MS equipment. The magnetic bead chemistries of IMAC-Cu, weak cation exchange (WCX) and strong anion exchange (SAX) were tested for suitability in sample fractionation.

An advantage with working closely with the experimenters is chip location of the samples and the run order are easily obtained. This allows diligent checking for batch effects, one of the primary concerns regarding bias in the TOF-MS system.

Asthma1

The spectra for the asthma1 dataset were obtained using IMAC-Cu magnetic beads. The plasma from each mother was sampled and divided into three sub-samples on which magnetic beads fractionation was performed separately (experimental replicates). Each of these sub-samples were sampled three times (technical replicates), each occupying a spot on the Bruker Daltronics Anchorchip MALDI target ($16 \times 24 = 384$ available spots), resulting in nine replicates per sample.

Unfortunately at the time of sample preparation, plasma samples of three mothers existed in insufficient quantity to be prepared for fractionation in triplicate. Two mothers were from the female birth group and one mother from the male birth group. Therefore, plasma from $n_{\mathrm{F}} = 13$ mothers with female births and $n_{\mathrm{M}} = 14$ mothers with male births was used. All $27 \times 9 = 243$ spectra were obtained in one run on the single MALDI chip with 117 and 126 spectra corresponding to the female and male birth groups, respectively.

Asthma2

From previous investigations by the APC group, birth-sex differentiating proteins with approximate weights of 9kDa were found. The WCX fractionation was used to complement the Murphy et al. (2005) study. As the 9kDa range was of interest the system was calibrated to produce spectra in the 1000 to $12000 m/z$ range.

Similarly to the asthma1 data, each plasma sample was fractionated in triplicate and subsequently each fractionation was spotted in triplicate on the MALDI chip, resulting in 270 spectra from the $n = 30$ mothers. Figure 1.8 illustrates the location and run order of samples on the MALDI chip.

From initial checking of the asthma2 data, it was apparent the spectra were particularly noisy and certain spectra needed to be excluded. Figure 1.9 was used to assess the variability and noisiness of the 270 spectra. This figure is a scatter plot of each spectrum's median absolute deviation (MAD, a metric of variability of a spectrum) against its initial log total ion current (a summation of the total amount of signal detected). Instead of points on the graphic, scaled spectra appear on the plot confined to equally sized rectangular areas, achieved by scaling spectra using their maximum intensity. A visible correlation between the variability of the spectra and the initial total ion current was seen. This correlation was used as a heuristic to help identify and remove unduly noisy spectra. After removal of 75 spectra (the noisiest spectra in the top-left corner of Figure 1.9), 96 female spectra and 99 male spectra remained corresponding to $n_{\mathrm{F}} = 14$ and $n_{\mathrm{M}} = 15$ mothers.

It was subsequently discovered that the samples for this dataset were not stored at sufficiently cool temperatures in transit. Samples were stored incorrectly at -20°C when they should have been stored at -80°C (the storage temperature at which they are kept in Melbourne). The APC observed protein peaks disappearing from technical replicates generated using the same fractionation chemistries in the preparatory MALDI TOF-MS runs which prompted investigation into the anomalies.

The asthma2 data are therefore not expected to have a true discriminatory signal between the experimental groups of newborn sex. However, the asthma2 dataset

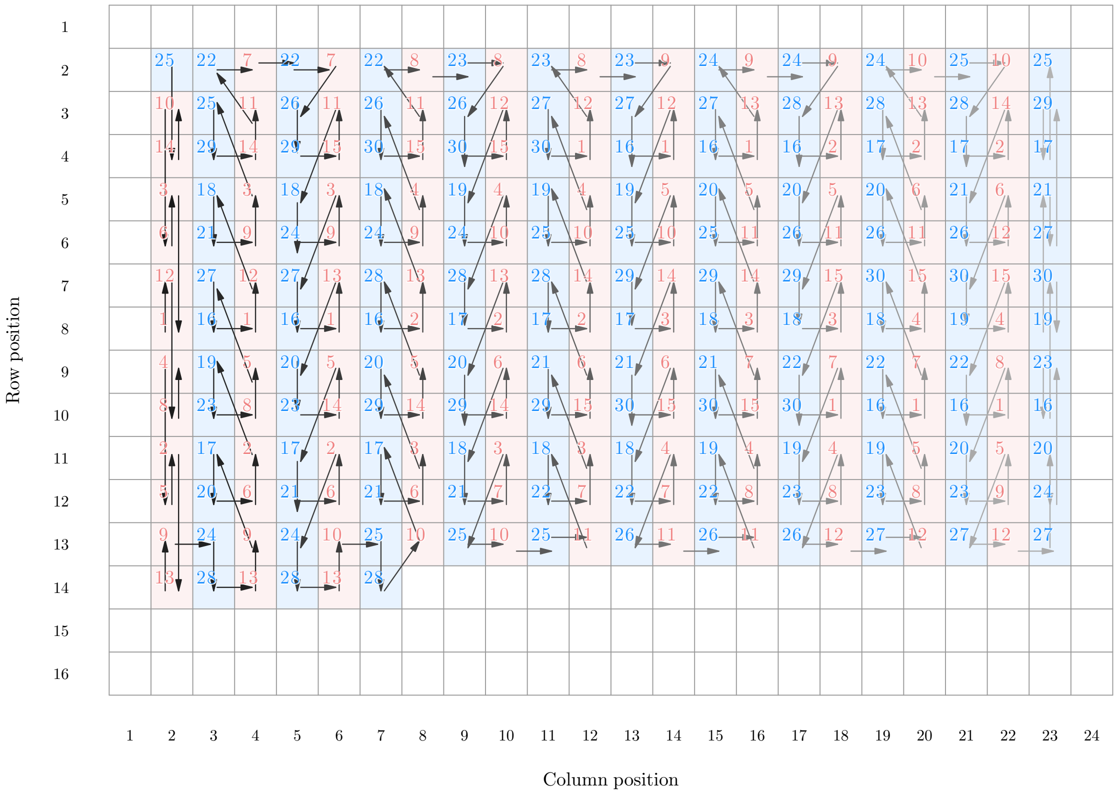


Figure 1.8: Schematic of location and run order of samples on the MALDI Bruker Daltronics Anchorchip for the asthma2 data. The numbers represent mother number, blue represents male foetus and coral female foetus. The arrows show the run/desorption order of the samples stating at location (2,2) and finishing at location (2,23) on the 16 × 24 spot chip. The run order generally follows a sequential progression of spots in 2 × 2 blocks as a fifth calibration spot exists (not pictured) in the centres that are irradiated prior to the 2 × 2 blocks for calibration.

will undergo the same analysis as the other datasets; it serves as a control as it is expected to provide null results.

1.3.6 Gastric cancer mice study

These data are related to published studies from the Ludwig Institute for Cancer Research (Tebbutt et al., 2002; Jenkins et al., 2005) and more recently in collaboration with the APC (Penno et al., 2012). Judd et al. (2009) also provides a good overview of the molecular mechanics at play. Ethics approval was obtained from the Ludwig Institute for Cancer Research Ethics Committee.

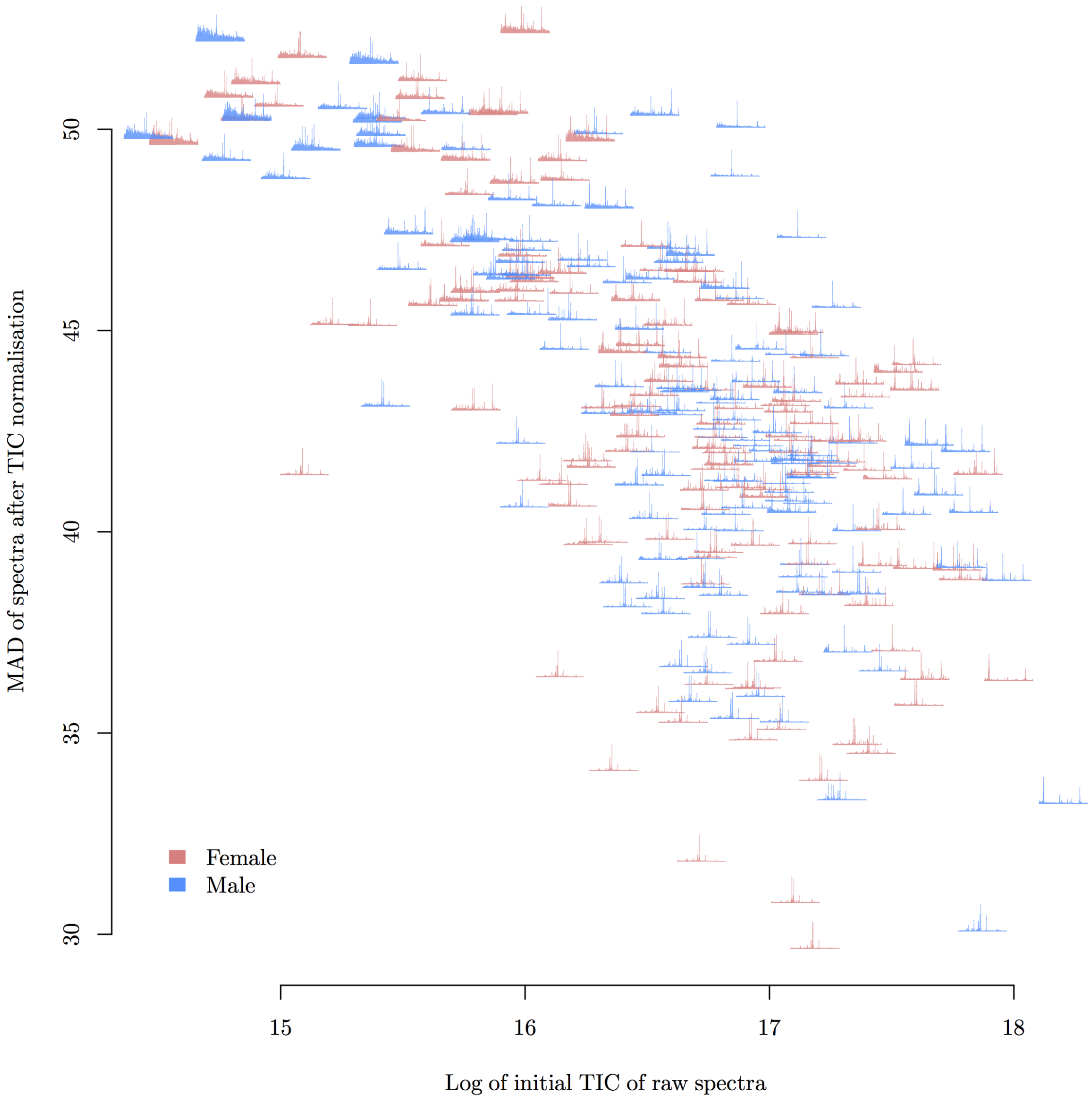


Figure 1.9: Scatter plot of the median absolute deviation of spectra against the natural logarithm of total signal detected for each spectrum in the asthma2 dataset. Instead of points on the graphic, a small visual of each spectrum itself is used, where each spectrum populates the same sized rectangle via scaling of intensities by the largest intensity. Spectrum colour indicates group classification.

Table 1.4: Groups in the GC mice data.

Group	Explanation	GC status	Inflammation	Total mice
WT	Control (wild type) mice	-	-	8
IL6	No IL6 gene (IL6$^{-/-}$) homozygotes for inflammatory suppression	-	-	8
FFStat3	Phe (F) mutation of gp130 but with Stat3 heterozygote (protective effect)	-	+	8
FFIL6	F mutation of gp130 with IL6 gene knockout homozygotes	+	-	8
FF	F mutation of gp130	+	+	8
				40

These data are serum-derived MALDI TOF-MS of mutated variants of mouse genotypes. There are two overarching groups of interest: gastric cancer (GC) and control phenotypes. These two groups can be further partitioned into five experimental groups in total, based on phenotype genetic variants (two GC groups, three control) as set out in Table 1.4.

The GC in the experimental mice is caused by a gene mutation in the glycoprotein 130 (gp130) protein coding gene. The mutation is a single amino acid difference (tyrosine, Y, to phenylalanine, F) in the translated gp130 cytokine protein, involved in inter-cell communication (signal transduction). The gp130 protein interacts with interlukin-6 (IL6) where the IL6 cytokine is involved in the inflammatory response of disease and foreign bodies. This mutated signal causes unhealthily high DNA transcription via JAK/STAT (janus kinase/signal transducer and activator of transcription) pathways leading to GC in the mice. The purpose of the sub-groups (seen in Table 1.4) within GC and control is to reduce confounding between the inflammatory response related with the IL6 group expression and the actual phenotype of malignant tumours. It is therefore hoped biomarkers that are found will not be confounded with inflammation. Although this is a murine model of GC, the IL6 gene is present in the human genome and it is hoped insight into early stage GC in humans can be obtained.

The blood samples of each of the $n = 40$ mice were taken at 12 weeks of age. Each sample was aliquoted into three subsamples where each subsample was allocated to one of the three MALDI chips. Within each aliquot, in a similar manner to Callesen et al. (2008), the subsample was split into three independent fractionations using C8 magnetic beads. Each of these fractionations were further sub-sampled to be allocated to three separate MALDI chip spots, totalling 27 spectra for each mouse

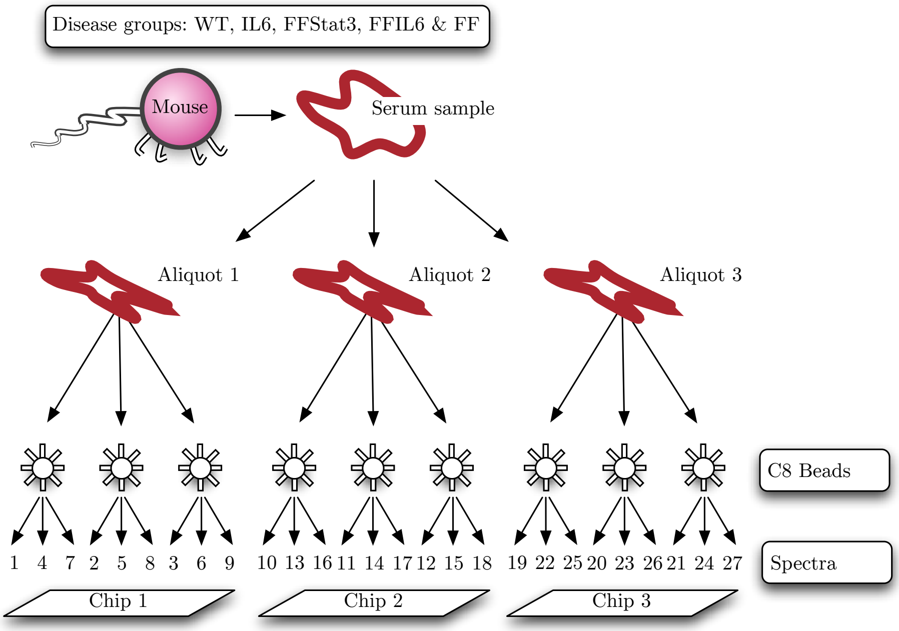


Figure 1.10: Experimental design for the GC mice data.

(nine on each chip). For greater clarity please refer to Figure 1.10 to see a schematic of the experimental design for *each mouse*.

These data, 1080 spectra in all, require different analysis to the other data outlined in this section because of the additional variables arising from the multi-chip experimental design. This adds complexity to the analysis but allows estimation of experimental biases that arise in the presence of spatial and temporal factors.

1.4 Summary of data

The Adam et al. (2002) data initially motivated this thesis. However, it has been established that the SELDI technology is largely not reproducible. The de Noo et al. (2006) dataset has been shown to provide strong classification signal. Greater than 90% correct classification was reported in both the de Noo et al. (2006) and Alexandrov et al. (2009) papers. Some design issues already discussed might be confounding the differentiation of the cancer and control subjects. The GC mice dataset is the flagship dataset of this thesis as it contains a large amount of replication and controlled experimental factors including a well defined murine-model. In addition, the effects of the hierarchical structure in these data can be investigated. The asthma datasets are not cancer-based data and are not expected to contain strong classification signal, in particular the asthma2 dataset.

Table 1.5 summarises the group abbreviations in these data. The colours of the groups are presented as they are kept constant throughout this thesis. In addition, Table 1.6 provides a summary of the MS data and relevant attributes for ease of reference.

Table 1.5: Summary of experimental groups for the datasets.

	Group		
Dataset	Description	Abbreviation	Colour
GC Mice†	Phe mutation	FF	🟧
	Phe mutation/IL6 knockout	FFIL6	🟨
	Phe mutation/Stat3 protective	FFStat3	🟦
	IL6 knockout	IL6	🟪
	Wild type control	WT	🟦
Adam et al. (2002)	Non-organ confined PC	CanB	🟪
	Organ confined PC	CanA	🟦
	Benign hyperplasia	BHyp	🟩
	Healthy control	Cont	🟧
de Noo et al. (2006)	Colorectal cancer	Canc	🟪
	Healthy control	Cont	🟦
Asthma1	Female birth	F	🟥
	Male birth	M	🟦
Asthma2	Female birth	F	🟥
	Male birth	M	🟦

†Please see Table 1.4 for detailed requisite information.

The remainder of this thesis is organised as follows. Chapters 2 and 3 assess current and novel pre-processing methods for their effectiveness. Upon successfully pre-processing the spectra, the peak expression data are available for analysis. Chapter 4 performs exploratory analysis and biomarker identification using linear models and new variance reduction techniques. Also considered is how missing values, often overlooked in MALDI/SELDI TOF-MS analysis, influence inference. Chapters 5 and 6 assess the signal in the datasets to differentiate experimental groups. Chapter 5 introduces the feature selection and classification models that are applied in Chapter 6. Chapter 6 explores some of the many practical data-generation options that potentially affect classification error. Please note that self-written and relevant computer code used throughout this thesis has been made available at `https://github.com/tystan/thesis`. References to the computer code within the text of this thesis are made to Appendix A where summaries of the computer code are available with links to specific portions of the code within `https://github.com/tystan/thesis`.

Table 1.6: Summary of important features of the MS datasets.

	Cancer/		Groups		Raw data	Number of m/z	
Study	disease	Technology	(K)	$n_1/n_2/\ldots/n_K$	m/z range	values $(T)^\dagger$	Comments
GC mice	Gastric	MALDI	5	8/8/8/8/8	600-20000	51000	Replicates
Adam et al. (2002)	Prostate	SELDI	4	81/78/84/83	0-200000	50000	
de Noo et al. (2006)	Colorectal	MALDI	2	48/64	960-11000	16000	
Asthma1	Asthma	MALDI	2	13/14	1996-40000	69000	Replicates
Asthma2	Asthma	MALDI	2	14/15	1000-12000	34000	Replicates
Morris et al. (2005)	NA††	SELDI	1††	100	935-33710	21000	Synthetic

†Approximate T.

†† These data are not created for discrimination. There are 100 spectra for each set of population parameters.

Chapter 2

Methods of intra-spectra pre-processing

The aim of creating the proteomic mass spectra is to analyse the profiles. Whether the primary aim of analysis is to identify biomarkers or to perform discrimination between groups, such analyses cannot be undertaken without removal of noise and systematic bias in the spectra. The description of this process, called data pre-processing, is separated into two chapters: intra- and inter-spectra pre-processing. This chapter is concerned with the adjustments required to be undertaken on individual spectra (intra-spectra pre-processing), prior to the adjustments to make spectra comparable (inter-spectra pre-processing), the latter being addressed in Chapter 3. The two intra-spectra pre-processing steps of signal smoothing and baseline correction are addressed here. For signal smoothing, the Savitzky-Golay method is compared to alternatives. Similarly, for baseline correction, an existing method called the top-hat operator is compared to standard methods. A novel extension of the top-hat operator to manage the intricacies of TOF-MS data is presented.

Proteomic mass spectra are not able to be meaningfully analysed without data pre-processing because of experimental noise. Figure 2.1 illustrates the systematic bias contained in TOF-MS data that ideally should be removed by pre-processing. A gold standard of pre-processing steps is not established, nor is there a consensus on the order in which the steps should be undertaken (Coombes et al., 2004). Take for example, three popular pre-processing packages: `MALDIquant` (Gibb and Strimmer, 2012), `PROcess` (Li, 2005) and `XCMS` (Smith et al., 2006). Not only do they provide different methods for each of the pre-processing steps, some provide more than one option for each step. It is a difficult problem as related by Coombes et al. (2004), "low-level processing of mass spectra involves a number of complicated steps that interact in complex ways".

Pre-processing of TOF-MS data can be sub-divided into two parts: intra- and inter-spectra pre-processing. This chapter outlines standard and potential new methods of intra-spectra pre-processing, after establishing the need for accurate pre-processing. Chapter 3 will focus on methods for inter-spectra pre-processing.

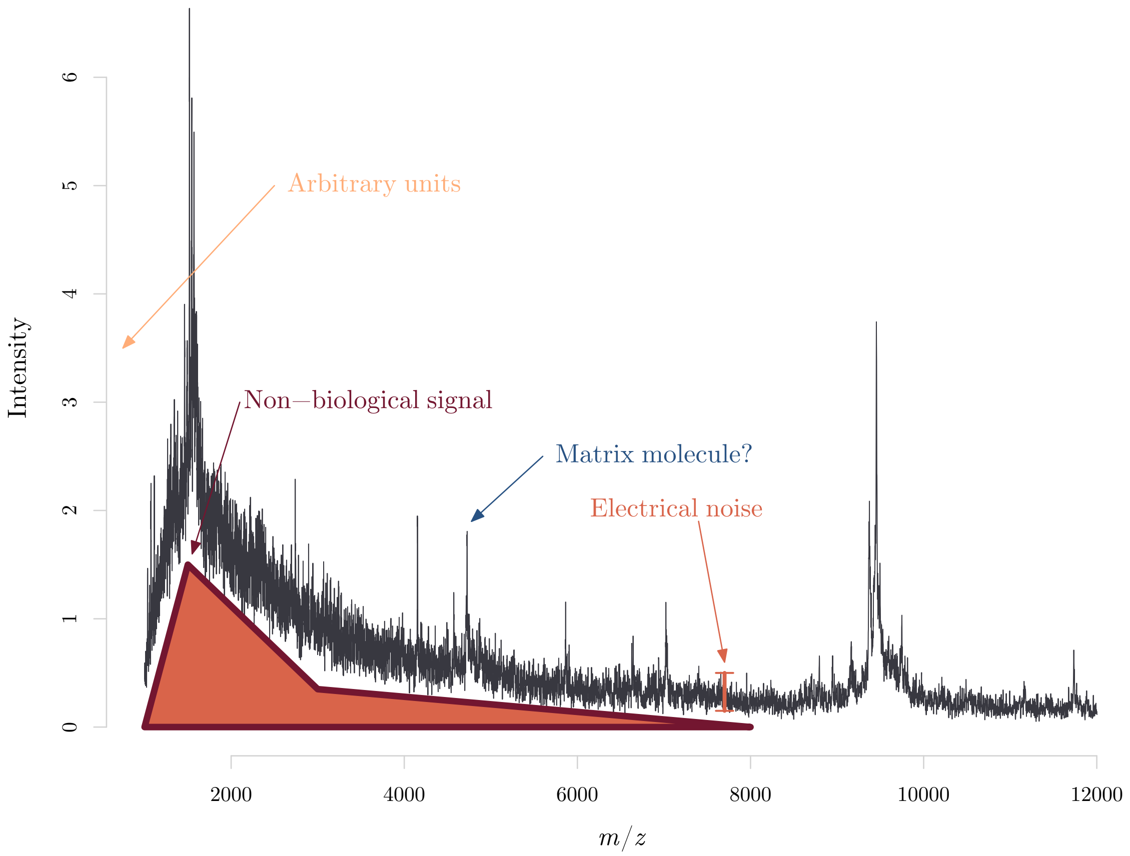


Figure 2.1: Sources of false signal in MALDI/SELDI TOF-MS data.

Figure 2.2 outlines the pipeline of pre-processing recommended here, to be carried out in the order shown. An explanation of the desired pre-processing order will follow an explanation of the purpose of the pre-processing steps.

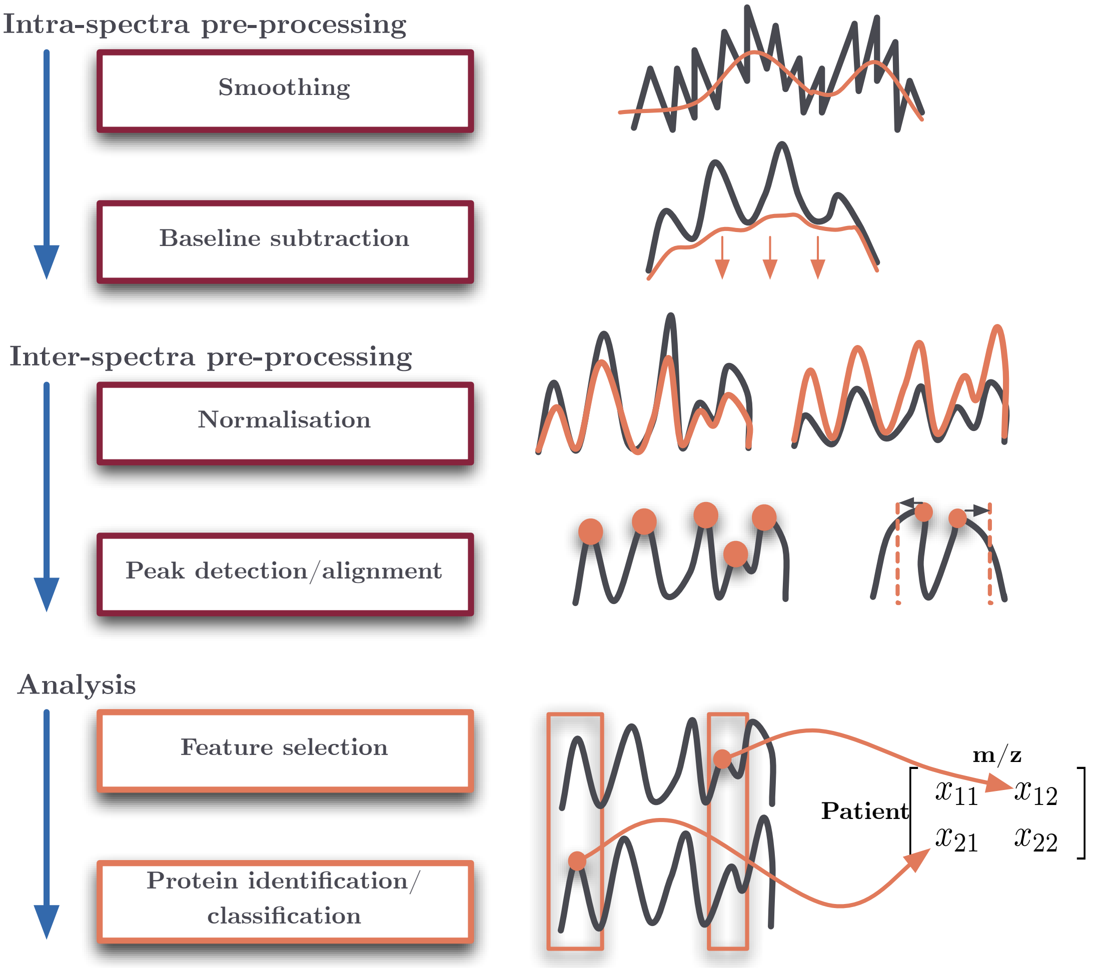


Figure 2.2: Pipeline of MALDI/SELDI TOF-MS pre-processing and analysis.

Signal smoothing is the first step in pre-processing the data: this is a denoising of the data to remove electrical oscillations present in the spectra signal. This precedes baseline subtraction which is the removal of additional, non-biological signal from ionised matrix particles and detector overload. These first steps in the process of eliminating false signal help enable observed peaks in the spectra to be a true representation of the intensity of charged peptides.

Normalisation is the first of the inter-spectra adjustments to make observed signals proportionate over the experiment, as instrument variability and sample ionisation will influence the number of charged peptides reaching the detector. Alignment of peaks, following peak detection, is required as there are small drifts in signal location by virtue of the calibration required for the TOF-MS system. From this pre-processed data, analysis to find differentially expressed peptides can be performed.

Smoothing should be performed before baseline subtraction as the electrical noise requiring removal by smoothing will erroneously affect baseline estimates. The baseline adjusted spectra should sit on a base of no signal. However, if smoothing is performed after baseline subtraction, the signal will sit incorrectly above zero intensity by virtue of smoothing non-zero, positive oscillating electrical noise.

Spectra must be normalised following baseline subtraction. Normalisation assumes the intensities are roughly proportionate across all spectra, so the inclusion of non-biological signal in normalisation will create incorrect adjustment of peaks across spectra.

Peak detection should be performed after normalisation, especially if non-global normalisation (§3.1) has been used. Many peak detection methods are a derivative of signal to noise calculations and will thus be affected by normalisation that adjusts spectra intensities non-uniformly. Peak alignment must occur post-peak detection, as the detected peaks are the entities of interest in alignment.

The pre-processing, along with visualisations and analysis, were performed in `R` (R Core Team, 2014). The `msProcess` package (Gong et al., 2012) was utilised for its data handling structures.[1] The majority of methods are performed using my self-written code; reference to use of existing packages and code will be made when applicable.

2.1 Pre-processing step I: signal smoothing

Different smoothing algorithms were considered for the smoothing pre-processing step. Savitzky-Golay (S-G; Savitzky and Golay, 1964) was identified as the leading smoother compared to the standard moving average smoothers.

S-G smoothing is a local regression method with the efficiency of pre-computed coefficients. Pre-computed coefficients are possible due to the local regression using a fixed-sized window of local points in each fit. This fixed-size window smoothing mimics a moving-average process but importantly is differentiated in its ability to preserve peak intensities irrespective of peak width (as long as the sliding local window is sufficiently wide).

[1] Archived from the CRAN repository on the 21/9/2012. The original maintainer was uncontactable (Lixin Gong). Used `msProcess` functions are written in `R`, visible and deemed reliable.

2.1.1 Savitzky-Golay smoothing

S-G smoothing considers distinct and evenly spaced time points, or m/z values, $t = 1, 2, \ldots, T$ and corresponding spectrum intensities $f_t = f(t)$. A least-squares polynomial of degree M in a local neighbourhood of each point t is fitted. That is, for each t and constants $L, R \in \mathbb{Z}^+$, the window of points $\{t-L, \ldots, t-1, t, t+1, \ldots, t+R\}$ is used in the local regression. However, as the time points are evenly spaced integers, a constant set of integers, $\{-L, \ldots, -1, 0, 1, \ldots, R\}$, can be used as the window of points for each $t = 1, 2, \ldots, T$.

The local linear regression assumes an M^{th} degree polynomial model at each point $t = 1, 2, \ldots, T$,

$$f_t\left(x\right) = \beta_{t0} + \beta_{t1}x + \beta_{t2}x^2 + \ldots + \beta_{tM}x^M, \tag{2.1}$$

for $x = -L, \ldots, -1, 0, 1, \ldots, R$. The S-G smoothed value at t,

$$f_{SG}\left(t\right) = \hat{f}_t\left(x=0\right) = \hat{\beta}_{t0},$$

is the local window least-squares estimate of $f(t)$. Therefore only the first element of the least squares estimate to Equation (2.1),

$$\hat{\boldsymbol{\beta}}_t = \left(X^T X\right)^{-1} X^T \boldsymbol{f}_t,$$

is required, where X is the design matrix of the constant point polynomial linear regression and $\boldsymbol{f}_t = \left(f_{t-L}, \ldots, f_{t-1}, f_t, f_{t+1}, \ldots, f_{t+R}\right)^T$. The S-G smoothed value $\hat{\beta}_{t0}$ for the point t can thus be expressed as $\hat{\beta}_{t0} = \boldsymbol{c}^T \boldsymbol{f}_t = \sum_{j=-L}^{R} c_j f_{t-j}$ where $\boldsymbol{c}^T = \left(X^T X\right)^{-1}_{[1]} X^T$ and $A_{[1]}$ denotes the first row of A. Thus, S-G smoothing is a moving-average filter with weightings,

$$\boldsymbol{c}^T = \left[\begin{array}{cccccccccc} c_{-L} & c_{-L+1} & \ldots & c_{-1} & c_0 & c_1 & \ldots & c_{R-1} & c_R \end{array}\right],$$

as $\boldsymbol{c}^T$ can be computed before the intensity values are observed (Press et al., 1992).

The weightings c_j are not necessarily positive values but sum to one, i.e. $\sum_{j=-L}^{R} c_j = 1$ (Orfanidis, 1996). The advantage of the S-G smoothing, than compared to a standard moving-average method, is that predictions of a local linear regression are used while maintaining the efficiency of moving-average filters. Moving-average coefficients are calculated or defined once, prior to any other computation and need not be recalculated.

The S-G method can be made more efficient by the fact that only the first element of the regression coefficient parameter estimates is required. Only the first row of $\left(X^T X\right)^{-1}$ is required to calculate the first element of $\hat{\boldsymbol{\beta}}_t$, where X is the pre-specified design matrix of S-G corresponding to Equation (2.1). This can be computed cheaply by using a lower-upper decomposition of X^TX (Press et al., 1992). i.e. $X^TX = LU$ where L is a lower triangular matrix and U is an upper triangular matrix. Therefore $\left(X^TX\right)^{-1} = U^{-1}L^{-1}$ and only the first row of U^{-1} is required.

An additional benefit of S-G is that peak heights tend to be preserved irrespective of the peak width. Traditional moving-averages tend to degrade narrowing peaks. For MALDI/SELDI TOF-MS, such considerations are extremely important as peaks widen for larger time values and is exaggerated more so on the m/z-scale.

2.1.2 Comparison to moving-average filters

Similar to S-G, a moving-average filter takes the form

$$m\left(t\right) = \sum_{j=-L}^{R} c_j f\left(t+j\right),$$

where $\sum_{j=-L}^{R} c_j = 1$ and $c_j > 0$.

A natural choice for the coefficients is
$\left[\begin{array}{cccccccc} c_{-L} & c_{-L+1} & \ldots & c_{-1} & c_0 & c_1 & \ldots & c_{R-1} & c_R \end{array}\right] =$

$$\left[\begin{array}{ccccccccc} \frac{1}{2(n_m-1)} & \frac{1}{n_m-1} & \ldots & \frac{1}{n_m-1} & \frac{1}{n_m-1} & \frac{1}{n_m-1} & \ldots & \frac{1}{n_m-1} & \frac{1}{2(n_m-1)} \end{array}\right],$$

where $n_m = L + R + 1$ with $L = R$ odd. This is the moving-average formulation used in comparison to S-G here. Figure 2.3 shows the two smoothing methods on a randomly chosen spectrum in the GC mice data. The degree of the polynomial and number of points used for the S-G smoothing are 4 and 51 respectively and the number of points used for the moving-average is 21. These were the input values that provided optimal smoothing as assessed by visual inspection. The prior exploratory analyses evaluated S-G smoothing with polynomial degrees $\{2,3,4\}$ and window sizes $\{25,51,75,101,125,151,175,201,301,601\}$, and moving-average smoothing using window sizes $\{5,11,15,21,25,31,41,51,75,101,201\}$.

From Figure 2.3 it is apparent that the moving-average filter does not preserve peak intensities to the same extent as S-G. Degradation of peak heights by the moving-average can be overcome by reducing the number of points used but has the result of retaining more noise. It can be seen the S-G filter is 'smoother' for the current calibration over the moving-average; if the number of points for the moving-average were reduced, the difference in 'smoothness' would become more pronounced. Another side-effect of reducing the points in the moving-average window is the creation of erroneous peaks from small sections of electrical noise.

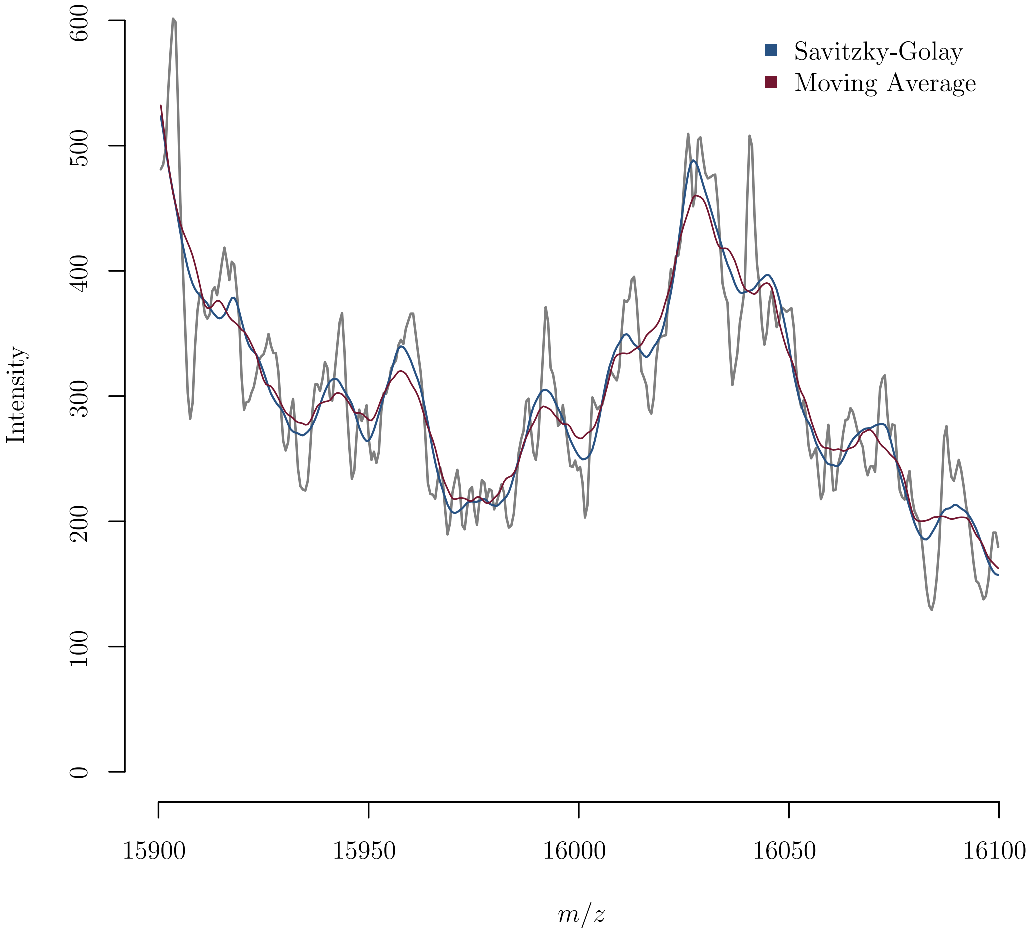


Figure 2.3: A randomly selected raw spectrum from the GC mice data on a subset of the m/z-axis with overlays of Savitzky-Golay and moving-average smoothing techniques.

2.1.3 Further considerations

Other considerations with smoothing were examined beyond the number of points used for the S-G and moving average windows.

(1) Additional smoothing of the residuals between the original signal and the smoothed signal can also be added back to the resultant smoothed signal. This caused little change in peak intensities but added spurious signals in relatively flat areas of the spectra.

(2) The literature recommends low, even-degree polynomials such as 2 or 4 (Bromba and Ziegler, 1981; Press et al., 1992) to be used in S-G smoothing. Polynomials of degree 4 allowed the flexibility to best smooth the curvature in the MS peaks than compared to degrees 2 and 3 in the exploratory analysis.

The code used for the S-G smoothing performed here is built on a flexible implementation available in the pracma R package (Borchers, 2012). The R package MALDIquant[2] includes S-G smoothing as its default spectrum denoising method, with default polynomial order of 3 and window size of 21 (half window size of 10). The alternate smoothing algorithm available in MALDIquant is the moving-average method with default window size of 5 (half window size of 2). The package documentation of MALDIquant points out that the window size for the moving average needs to be much smaller than for S-G in most cases, an observation additionally made here. The Bioconductor packages XCMS (Smith et al., 2006) and MassSpecWavelet (Du et al., 2006) contain S-G functionality, however the XCMS package is specifically designed for liquid chromatography MS and the MassSpecWavelet package is specific to wavelet methods to detect peaks only.

2.2 Pre-processing step II: baseline correction

The method of baseline correction is the second of four major pre-processing steps in the reduction of observed intensities to remove extra (false) signals to allow meaningful analysis of MS data. Once the electrical noise is removed from proteomic MALDI-TOF MS signals via smoothing, baseline correction needs to be applied. The additional, non-biological signal removed by baseline subtraction is a result of overload of the TOF detector (especially at the low TOF-values which approximately equate to low proteomic mass values) and non-peptide, ionised matrix particles in the TOF-system. Baseline correction is a way of eliminating false signal and thus aims to leave peaks that are true expressions of charged peptides and not other artefacts.

The baseline subtraction method proposed here is called the *top-hat* operator that uses *morphological openings*. The top-hat operator is a non-parametric, non-linear filter. Its advantages over standard methods will be explored in the following sections.

Mathematical morphology was initially proposed in two-dimensional image analysis prior to modern ‘omics’ analysis but is now used in the analysis of gene expression data (Yang et al., 2002; Mayer and Glasbey, 2005). A one-dimensional form of this operator is successfully applied here to TOF-MS data extending the baseline correction method from Sauve and Speed (2004).

Morphological openings have desirable properties for baseline correction. For example, the false signal in MS spectra may not have a known functional form, and the morphological opening assumes none. Additionally, the calculation of morphologi-

[2] Version 1.7 released May 2013 during writing of this thesis first included S-G smoothing.

cal openings is computationally inexpensive in comparison to some functional filters that require estimates of model parameters. Before elaborating on the details of the advantages of this method, an overview of morphological theory with definitions developed in the image analysis area is presented in §2.2.1.

2.2.1 Morphological image analysis and theory

To begin, a formal definition of a supremum is given below. A supremum for a subset, S in $\mathbb{R}$ ($S \subset \mathbb{R}$) can be written as follows.

Definition 2.1: Supremum (Bauldry, 2009). *The supremum is a number, r, where*

(1) r is an upper bound for S ($s \leq r \; \forall \; s \in S$).

(2) If R is another upper bound of S, then $r \leq R$.

For the purposes of MS data, $\sup(S)$ ('supremum of S'), is the maximum value in S. The formal definition of a supremum allows r to be a number that may not necessarily be part of the set, S. Similarly, an infimum (inf) is the point-wise minimum of some real values.

Further, the following equalities hold (Bauldry, 2009),

$$\begin{aligned} \sup(S) &= -\inf(-S), \text{ and} \\ \inf(S) &= -\sup(-S). \end{aligned}$$

The core concepts of morphological image analysis are presented below to make the discussions that follow self-contained.

Definition 2.2: Structuring element (Soille, 1999). *A structuring element (SE) is a small set that acts on given data/images.*

Definition 2.3: Centred SE (Soille, 1999). *A centred SE is a set where the median value is 0.*

There are two types of SEs: flat SEs and non-flat SEs. Flat SEs are small sets of elements of the same dimension as the data or image. For example, with regard to TOF-MS data, a flat SE is simply a one-dimensional window passed over the one-dimensional vector of spectral intensities. Non-flat SEs are SEs one-dimension higher than the input data. For example, a non-flat SE for mass spectra would be a flat SE but with weights assigned to positions in the set according to location of the elements in the SE.

The definitions of the morphological operators: *dilation*, *erosion*, *opening*, *closing* and *top-hat* are presented below for flat and non-flat SEs for completeness and understanding. However, the focus in this chapter will be on flat SEs as this is the SE required for successful baseline correction. Additionally, flat SEs will be assumed to be centred, symmetric, closed sets. That is, the SE will behave the same on the data on the left of the SE's centre as it does on the right.

Definition 2.4: Erosion with a non-flat SE (Soille, 1999). *For the sets $X \subset \mathbb{Z}^p$ and $B \subset \mathbb{Z}^{p+1}$, $p \in \mathbb{Z}^+$, and the functions f and g, defined over X and B respectively, the erosion of X by B is defined as,*

$$\begin{aligned} \epsilon_g(f)(x) &:= (f \ominus g)(x) \\ &:= \inf_{b \in B} f(x+b) - g(b). \end{aligned}$$

Definition 2.5: Erosion with a flat SE (Soille, 1999). *Once again, for the set $X \subset \mathbb{Z}^p$ but now $B \subset \mathbb{Z}^p$, $p \in \mathbb{Z}^+$, and the function f defined over X, the erosion of X by B is defined as,*

$$\begin{aligned} \epsilon_B(f)(x) &:= (f \ominus B)(x) \\ &:= \inf_{b \in B} f(x+b). \end{aligned}$$

Definitions 2.4 and 2.5 can be interpreted where $f(x)$ is the object of interest (e.g. mass spectrum intensities), $x \in X$ are the indexes of the $f(x)$ intensities and B is the SE. The function g in the non-flat case is simply a weight function as the SE passes over the set of interest, $f(x)$, $x \in X$. This weight function adds an extra dimension to the SE as per the definitions of flat and non-flat SEs.

Definition 2.6: Dilation with a non-flat SE (Soille, 1999). *For the sets $X \subset \mathbb{Z}^p$ and $B \subset \mathbb{Z}^{p+1}$, $p \in \mathbb{Z}^+$, and the functions f and g, defined over X and B respectively, the dilation of X by B is defined as,*

$$\begin{aligned} \delta_g(f)(x) &:= (f \oplus g)(x) \\ &:= \sup_{b \in B} f(x+b) + g(b). \end{aligned}$$

Definition 2.7: Dilation with a flat SE (Soille, 1999). *Once again, for the set $X \subset \mathbb{Z}^p$ but now $B \subset \mathbb{Z}^p$, $p \in \mathbb{Z}^+$, and the function f defined over X, the dilation of X by B is defined as,*

$$\begin{aligned} \delta_B(f)(x) &:= (f \oplus B)(x) \\ &:= \sup_{b \in B} f(x+b). \end{aligned}$$

Dilation is the dual operator of erosion and simply finds the maximal value of $f(x)$ in a domain defined by the SE, B, and uses the supremum as opposed to infimum.

Definition 2.8: Morphological opening and closing (Soille, 1999). *For a flat SE, B, and a set X, the opening of X by B is defined as,*

$$\begin{aligned}\omega_B(f)(x) &:= \delta_B(\epsilon_B(f))(x)\\ &:= (f \ominus B) \oplus B.\end{aligned}$$

Similarly, a closing is defined as

$$\begin{aligned}\psi_B(f)(x) &:= \epsilon_B(\delta_B(f))(x)\\ &:= (f \oplus B) \ominus B.\end{aligned}$$

Note, the definition for an opening and closing for a non-flat SE, ω_g and ψ_g respectively, are defined similarly.

In this context, a morphological opening is a non-linear filter that estimates a background signal of the one-dimensional spectrum X. The opening has the property that

$$\omega_B \leq f \quad \forall\, x \in X.$$

Similarly, the closing has the property

$$\psi_B \geq f \quad \forall\, x \in X.$$

Definition 2.9: Top-hat operator (Soille, 1999). *The top-hat operator is the residual of the set X from the opening of X for a defined SE, B. For a flat SE, the top-hat operator is defined as,*

$$\tau_B(f)(x) := f(x) - \omega_B(f)(x).$$

Of course, the top-hat operator for a non-flat SE is similarly defined as $\tau_g(f)(x)$.

In other words, the result of the top-hat operator is the (non-linear) estimation of the true signal by removing the background signal from X. Because of the $\omega_B(f) \leq f$ $(\forall\, x)$ property of morphological openings, the top-hat operator provides a conservative background adjustment and removal without risk of creating negative signal which is a physical impossibility of the system.

Example

To illustrate morphological operators, consider a simple example. Let $f = \{a_x\}_{x=1}^{13}$ be a series and define a flat SE, $B = \{b_j\}_{j=1}^{5} = \{-2, -1, 0, 1, 2\}$ with

$$f(x) = \begin{cases} a_1 & \text{if } x < 1 \\ a_x & \text{if } x = 1, 2, \ldots, 13 \\ a_{13} & \text{if } x > 13 \end{cases}$$

where

$$\{a_x\} = \left\{ \begin{array}{ccccccccccccc} 6 & 11 & 12 & 14 & 7 & 10 & 13 & 9 & 12 & 15 & 8 & 11 & 10 \end{array} \right\}.$$

Using the flat SE, $B = \{-2, -1, 0, 1, 2\}$, the effect of ϵ_B, ω_B and τ_B can be observed in Figure 2.4.

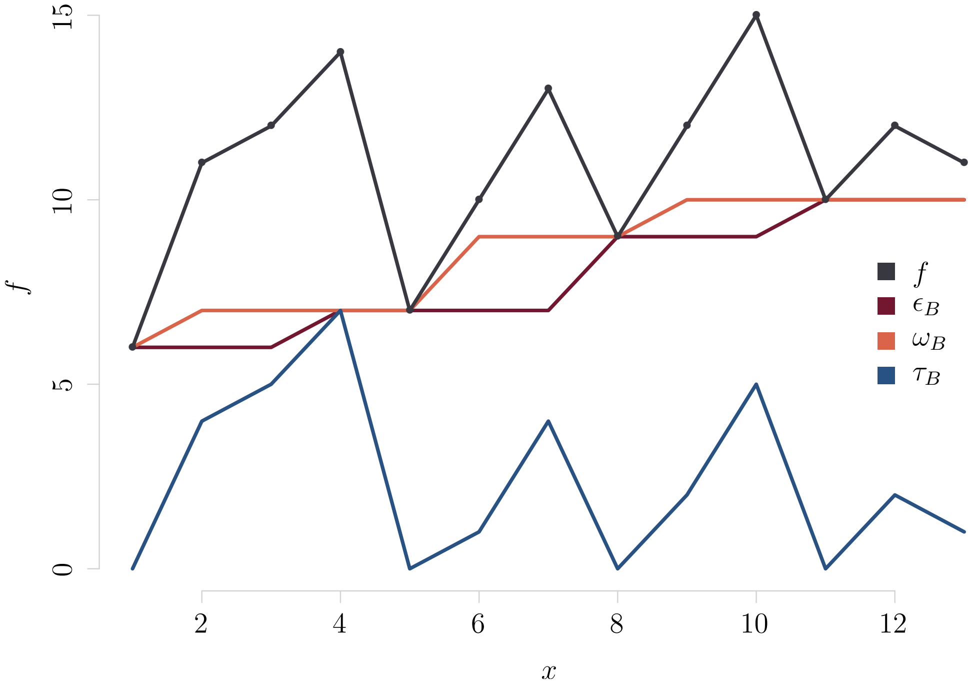


Figure 2.4: An example of ϵ_B, ω_B and τ_B on a set f.

2.2.2 Implementation

An erosion of a one-dimensional spectrum's expressions, f at evenly spaced points $x_1, x_2, \ldots, x_n$, is calculated using a moving window that traverses each of the x_i points, assigning the minimum value of f in the window to that point. A naive R implementation is available in Appendix A.1.

The SE needs to be chosen carefully.

(1) If a SE is too large then it will be too conservative and leave false signal.

(2) If a SE is too small will result in under-cut peaks and the removal of valid signal.

(3) The mean peak width gets larger further along the x-axis: the baseline subtraction needs to be performed piecewise otherwise issues (1) and (2) will occur.

Figure 2.5 presents a comparison of the top-hat operator with other standard methods of baseline subtraction. These standard methods are estimated by calculating local minima (troughs) and fitting either local regression (LOESS) or interpolating (splines) through these points (Yang et al., 2009). These standard methods require careful selection of window size for detecting troughs, polynomial order and the span of points for fitting the model where applicable. Despite using optimised input parameters for the standard methods, they cannot guarantee non-negative signal. In some cases, the standard methods presented may produce padded or removed signal in places of high curvature in the spectra. An example of this is in the 9750-10250m/z range in Figure 2.5.

An important consideration in the application of the baseline subtraction methods shown in Figure 2.5 is that they all need to be performed on subsets of the spectra with different input parameters. This piecewise approach is required because, on average, peaks become wider for larger m/z- or TOF-values (Zhang et al., 2010) and the baseline subtraction methods are sensitive to considerable changes in peak width.

The top-hat operator, when used piecewise for baseline subtraction for TOF-MS has many advantages over standard methods. A flexible estimation of spectra baseline is calculated as no functional form is assumed. The top-hat operator importantly preserves small biological signal as no negative signal will be created by the top-hat operator, τ_B. Additionally this baseline subtraction method is computationally efficient as no model parameters need be estimated.

Not only is the top-hat operator efficient, but it can also be significantly improved in performance by the use of the line-segment algorithm (§2.2.3). In the case of mass spectrometry, there may be tens of thousands of data points over thousands of spectra. For the naive looping erosion algorithm (as well as dilation and top-hat) presented in Appendix A.1, computation is unnecessarily inefficient. The obvious improvement on this algorithm is to remove repeated minimum operations on data points as the algorithm moves through the f series. The next section outlines a faster implementation of the top-hat algorithm largely overlooked to date by the mass spectrometry literature. Additionally, freely available MS software does not

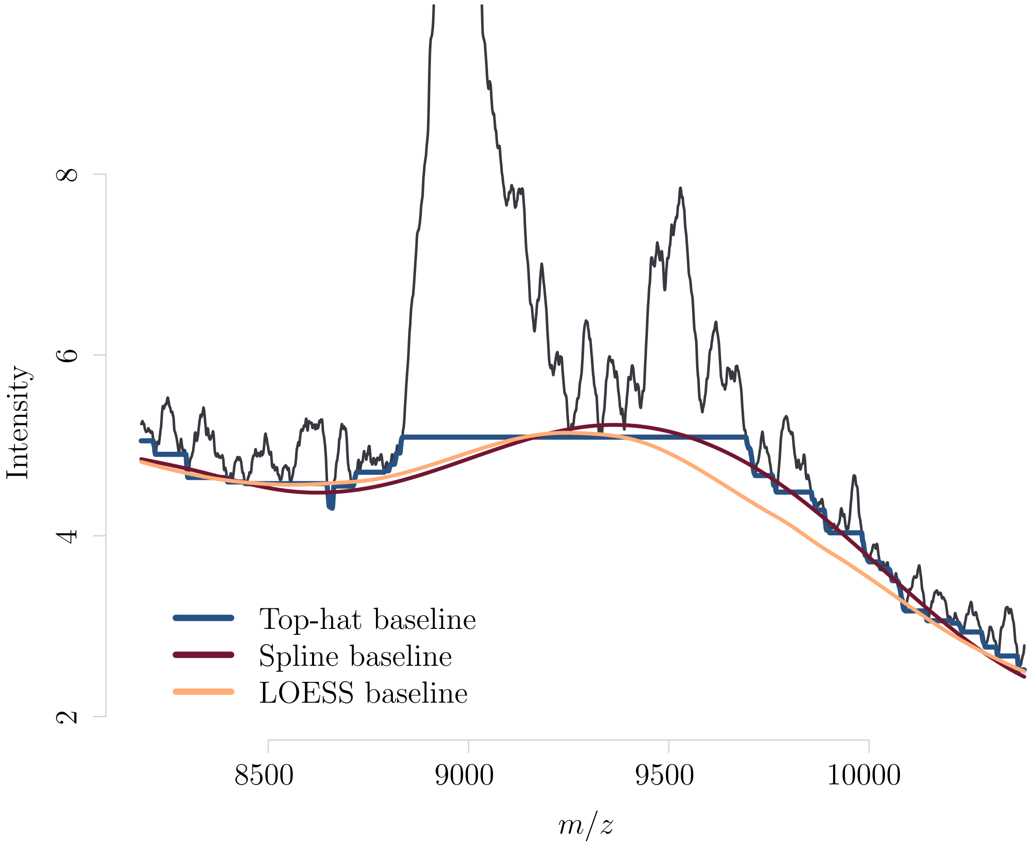


Figure 2.5: A spectrum from the asthma1 dataset demonstrating the baseline estimates in the signal using the top-hat operator, spline and LOESS methods.

implement morphological openings,[3] or if the top-hat operator is implemented, the faster algorithm is not.[4,5]

2.2.3 A more efficient implementation

A more efficient algorithm in computing erosions and dilations has been proposed by van Herk (1992) and also Gil and Werman (1993). Named the line segment algorithm (LSA), it has also been generalised to non-centred flat SEs by Soille (1999). Its application is mainly seen in medical imaging and analysis (van Herk et al., 1998; Heneghan et al., 2002).

[3] `Bioconductor` mass spectrometry packages or `CRAN-R`.

[4] As of November 2013 OpenMS/TOPP 1.11.1.

[5] With the exception of the `R` package `MALDIquant` where version 1.6 released March 2013 during writing of this thesis included top-hat baseline subtraction.

While the standard naive iterative algorithm will have k (the length of the SE) comparisons for each element of the input vector (as the SE will encompass each data point k times), the LSA requires only three comparisons per element irrespective of SE size.

The line segment algorithm (van Herk, 1992)

Calculation of moving window minimums (erosions) for the data,

$$X = \{1, 2, \dots, n\} \quad \text{and} \quad f(x) \text{ defined for } x \in X,$$

and a flat SE of length k ($\in \mathbb{Z}^+$, odd) centred at $k_0 = \frac{k+1}{2}$.

In this algorithm it is assumed that n is a multiple of k, i.e. $mk = n$, $m \in \mathbb{Z}^+$. Two temporary vectors of length n to finally compute $\epsilon_B(f)$ are used. The two temporary vectors are calculated as follows,

$$g(x) = \begin{cases} f(x) & \text{if } x = 1, k+1, 2k+1, \dots, (m-1)k+1 \\ \min\left[g(x-1), f(x)\right] & \text{otherwise.} \end{cases}$$

That is, g is created in one pass, sweeping from left to right. Similarly h is created from right to left,

$$h(x) = \begin{cases} f(x) & \text{if } x = mk, (m-1)k, (m-2)k, \dots, k \\ \min\left[f(x), h(x+1)\right] & \text{otherwise.} \end{cases}$$

The erosion of $f(x)$, $\epsilon_B(f)$, is found by comparing the temporary vectors g, h by,

$$\epsilon_B(f)(x) = \min\left[g(x+k_0), h(x-k_0)\right].$$

The stringent assumption that the length of X, n, is a multiple of the SE length, k, can be easily overcome. For cases where $mk \neq n$ for erosions, simply extend the length of the input vector f to the next multiple of k so the equality $mk = n$ holds. In the newly created elements, place a suitably large number or computational infinity. Then the algorithm can be applied and subsequently reduce the resulting vector to the original size by removing the inserted elements. The resulting erosion with this modification will be correct as the computational infinities added to the series will not change the results of the minimum calculations otherwise made in the algorithm.

As discussed in §2.2.1, $\epsilon_B(f)$ and $\delta_B(f)$ are dual operators, so the dilation algorithm is a matter of either:

(1) swapping *min* in the line segment algorithm for erosions with *max* for dilations, or,

(2) performing the erosion on $-f$ and returning the negative erosion to get the dilation. i.e. $\epsilon_B(f) = -\delta_B(-f)$.

If option (1) from the above is chosen, extending the length of f so $mk = n$, requires the insertion of negative computational infinity in f.

Finally, note that some values of $x \pm k_0$ are not elements of $\{1, 2, \ldots, n\}$, so no minimum calculation is required for $\epsilon_B(f)(x)$, where $g(x + k_0)$ is defined but not $h(x - k_0)$ and vice-versa.

The more efficient LSA (van Herk, 1992; Gil and Werman, 1993), requires only three comparisons per element irrespective of SE size. This converts the complexity of the computation of an erosion or dilation from $\mathcal{O}(kn)$ to $\mathcal{O}(n)$. An implementation of the LSA can be found in Appendix A.1.

To illustrate the increase in speed this algorithm provides on modern MALDI-TOF MS data pre-processing, a synthetic dataset of 200 spectra with randomly generated positive values of 50,000 data points each were created. Using a SE of size 301, a comparison of the time taken to baseline correct these data in practice can be seen in the Table 2.1. Not only is there an increase in speed in calculating the top-hat operator using the LSA but the increase in speed can be obtained by at least an order of magnitude using compiled C-code.

Table 2.1: Computation time of top-hat operator using different code on generated data.

Top-hat method	Time taken (seconds)†	Relative time
Standard (R-code)	251.8	49.4
Line-segment (R-code)	143.2	28.1
Standard (R using compiled C-code)	18.2	3.6
Line-segment (R using compiled C-code)	5.1	1††

†MacBook Pro7 (Intel Core 2 Duo 2.4 GHz, 3 MB L2 Cache, 8 GB Memory)

††Reference

2.2.4 Towards automated baseline correction

Despite the increase in speed in the calculation of morphological openings provided by the line segment algorithm, piecewise baseline correction is still required. The SE size used needs to be of equivalent window size to the spectra's peak widths, or greater, to ensure the top-hat operator does not undercut peak intensities. The piecewise baseline correction involves determining subsections of the m/z-axis where fixed SE widths in each section are appropriate, or the equivalent parameters for other baseline methods. Smaller SEs will be chosen for the lower m/z-values and larger SEs will be used for larger m/z-values.

In this section a novel method of fast baseline correction is proposed that requires no user input and is calibrated using pre-baseline corrected MS data.

To illustrate the complication of increasing average peak width across the x-axis, Figure 2.6 shows the relationship between the peak width and peak location for a subset of the GC mice dataset. Proteins have isotopic distributions as a result of naturally occurring isotopes in nature. For example, carbon-13 is a naturally occurring isotope (1.11% of carbon atoms; Zhang et al., 2010) found in organic molecules. The MS system cannot resolve isotopic distributions of proteins as individual peaks and thus a single peak shape is observed. The isotopic distribution is wider for larger mass proteins as a result of the increase of possible isotopic combinations. Proteins also exist with isoform variations that provide mass variation to the most common form of the protein. These isoforms can create peak broadening as the MS system cannot resolve the individual isoform peaks or be observed as separate peaks. For example, glycoproteins may have isoforms of the protein with different carbohydrates attached. The presence of protein isoforms can also be biological signals of interest (Pan et al., 2005).

Figure 2.6 is obtained after the pre-processing steps of smoothing, baseline correction and normalisation. However, rough estimates of peak widths based on distances between troughs (calculations outlined in §3.2) in the raw or smoothed spectra are sufficient to estimate the SE lengths required to automate the baseline correction proposed here. The estimated SE sizes will be assumed from here in and are used to create the transformation of the data so a constant sized SE can be applied to the transformed data. The morphological algorithms discussed up to this point cannot handle the transformed data but proposed here is an extension of LSA for data where the points are not equally spaced.

If a transformation of the TOF-axis can be made so the peak widths are generally constant across the transformed axis, then the piecewise approach is not required. Siuzdak (2006) and House et al. (2011) have suggested peak width is roughly proportional to peak location on the TOF-axis. This was not the case for the data analysed

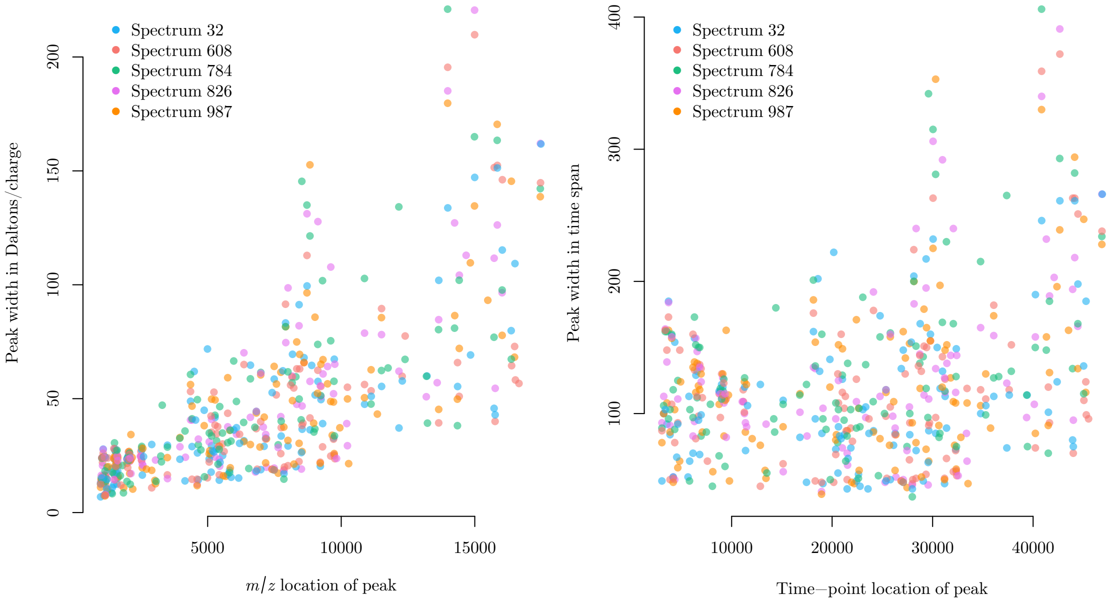


Figure 2.6: A random selection of five spectra from the GC mice dataset and their respective peak widths of detected peaks (see §3.2) at x-axis locations in terms of (a) m/z-value and (b) TOF-value.

in this thesis. Figure 2.7 shows a potential transformation of the TOF-axis which creates a roughly constant peak width irrespective of the transformed peak location. The thick horizontal grey line in Figure 2.7(b) represents a peak width/SE size that encompasses 97.5% of detected peaks, yielding successful baseline subtraction. Of course, a larger percentage could be used to ensure fewer peaks are ‘undercut’. Such considerations do not detract from the validity of the method as selecting SE size by inspection is unlikely to consider all the spectra. Additionally, as variability of peak widths exist with respect to the average width for a m/z-value, there is not a SE size that is optimal for all peaks across spectra at a given m/z value. Using different SE sizes for individual spectra is possible via the method outlined here but may induce bias as the adjustments to each spectrum are not consistent.

From Figure 2.7 it can be seen if an alternatively formulated top-hat operator that can be applied to the transformed, non-evenly spaced values of the TOF-axis for a constant SE width, the baseline subtraction step could be automated. Required changes to the morphological top-hat operator to work in this setting of non-constant spaced, non-integer x-values is addressed in §2.2.5. A novel LSA extension is presented in §2.2.6 for the case of non-equally spaced x-values, and finally, §2.2.7 compares the effectiveness of the piecewise and the newly formulated top-hat baseline subtraction methods.

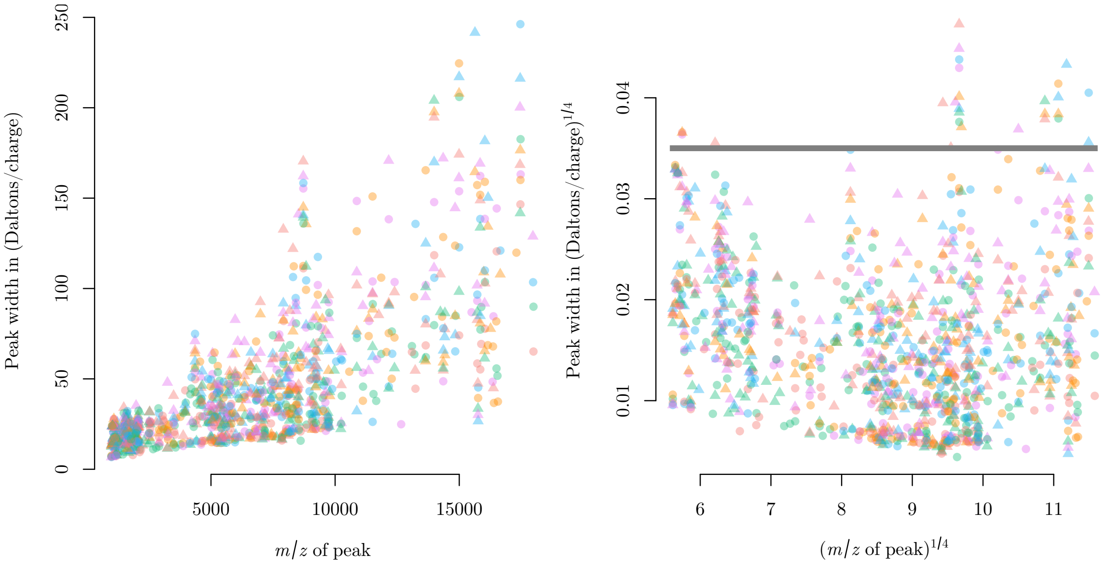


Figure 2.7: A random selection of ten spectra represented by different colour and symbol combinations from the GC mice dataset and their respective peak widths of detected peaks (see §3.2) at x-axis locations in terms of (a) m/z-value and (b) transformed TOF-value (TOF$^{1/4}$).

2.2.5 Morphological analysis for unequally spaced values

This section extends the top-hat operator outlined in §2.2.1 and §2.2.2 for creating a baseline subtraction method that minimises the need for user input. Automated methods of pre-processing are desirable to minimise user error and the time required to undertake pre-processing of the data. The novel algorithm proposed for handling MS data outlined in this section may have further applications to other areas for reducing computation time of moving window filters where data cannot be assumed equally spaced. For example, where a signal has been sampled at uneven intervals.

Consider the case where values in X are not evenly spaced, $X \subset \mathbb{R}$, rather than $X = \{1, 2, \ldots, n\} \subset \mathbb{Z}$, such as proteomic spectra on a transformed TOF-axis.

Definition 2.10: Erosion on arbitrarily spaced data. *Consider a set* $X = \{x_1, x_2, \ldots, x_n\}$ *with* $x_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$ *and* $x_i < x_j \quad \forall i < j$. *The function* $f : \mathbb{R} \to \mathbb{R}$ *is defined for all elements in the set* X. *The erosion of a mass spectrum* f *using a flat one-dimensional SE of size* k *over* X *is*

$$\epsilon_B(f)(x) = \begin{cases} \min\{f(x_t)\} \ \forall x_t \ s.t. \ x_1 \leq x_t \leq x + k_0 & if\ x \leq x_1 + k_0 \\ \min\{f(x_t)\} \ \forall x_t \ s.t. \ x - k_0 \leq x_t \leq x + k_0 & if\ x_1 + k_0 < x < x_n - k_0 \\ \min\{f(x_t)\} \ \forall x_t \ s.t. \ x - k_0 \leq x_t \leq x_n & if\ x \geq x_n - k_0. \end{cases}$$

In Figure 2.8, the calculation of $\epsilon_B(f)(x)$ requires each value $x \in X$ to be considered and all values in $X \subset \mathbb{R}$ that are within $k/2$ need to be found. The minimum value of all the respective mappings of these values by f is the erosion at that point. Appendix A.1 provides a naive implementation to obtain $\epsilon_B(f)(x)$ as per Definition 2.10.

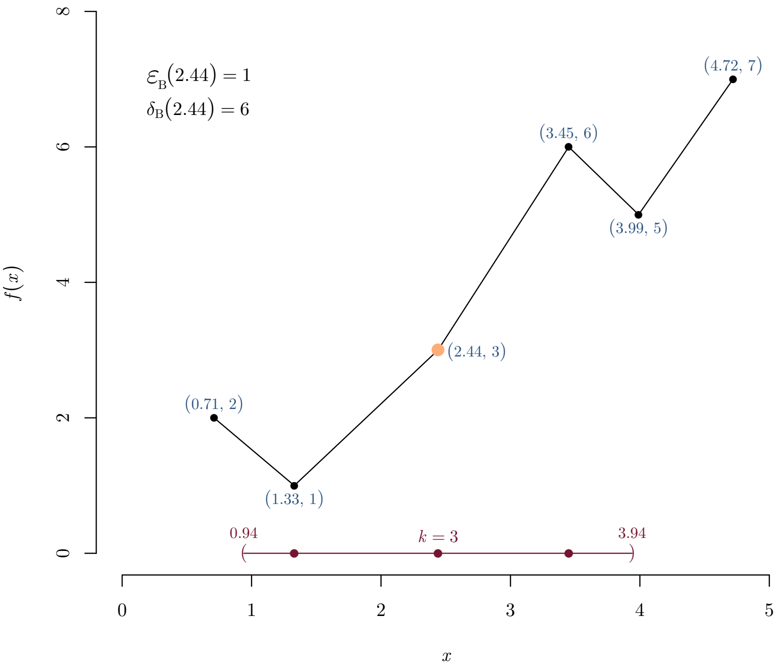


Figure 2.8: An example of data where x values are not evenly spaced, but rather, unevenly spaced points in $\mathbb{R}$. For a SE of $k = 3$ and the point $x = 2.44$, using Definition 2.10, it can be seen the erosion and dilation of this point are 1 and 6, respectively.

By considering a naive implementation of Definition 2.10, it can be observed the computational complexity of the algorithm is not optimised as calculations of minimum values are repeatedly and redundantly being performed on the same values, as was the case when the naive algorithm for discrete and evenly spaced values on the x-axis was compared to the LSA. In the next section, a novel algorithm to find minimum or maximum values for a sliding window will be given, named here the continuous line segment algorithm (CLSA).

2.2.6 The novel continuous line segment algorithm

Much of the image analysis literature focuses on evenly spaced data points (i.e. pixels), but in previous sections the need for morphological operators for unequally spaced data has been outlined. A new algorithm is proposed here that extends the

LSA for the continuous case, i.e. $X \subset \mathbb{R}$. As the elements in X are not evenly spaced, different strategies are required than those used in the LSA to compute moving window minimums or maximums. The CLSA is presented below.

The novel CLSA

Let $k_0 = \frac{k}{2}$ where k is the length of a centred, one-dimensional window. Consider the ordered (ascending) set of arbitrarily spaced points on the x-axis, $X = \{x_1, x_2, \ldots, x_n\}$, and the corresponding expression values, f. Furthermore, define $\text{span}(X) = x_n - x_1$ where $\text{span}(X) > k$ and choose the smallest $m \in \mathbb{Z}^+$ so that $mk \geq \text{span}(X)$.

Three vectors taking integer values are required to be created initially,

$$\Theta = [\theta_1, \theta_2, \ldots, \theta_n], W^{\triangledown} = [w_1^{\triangledown}, w_2^{\triangledown}, \ldots, w_n^{\triangledown}], W^{\vartriangle} = [w_1^{\vartriangle}, w_2^{\vartriangle}, \ldots, w_n^{\vartriangle}].$$

For $i = 1, 2, \ldots, n$, the integer θ_i is calculated as follows,

$$\theta_i = \{j : \text{if } x_1 + (j-1)k \leq x_i < x_1 + jk\} \text{ for } j \in \{1, 2, \ldots, m\};$$

$w_i^{\triangledown}$ is the index corresponding to x_i satisfying the inequality,

$$x_{w_i^{\triangledown}-1} < x_i - k_0 \leq x_{w_i^{\triangledown}};$$

and $w_i^{\vartriangle}$ is the index corresponding to x_i satisfying the inequality,

$$x_{w_i^{\vartriangle}} \leq x_i + k_0 < x_{w_i^{\vartriangle}+1}.$$

The moving window minimum at x_i can be calculated as,

$$r_{\text{CLSA}}\left(f\left(x_i\right)\right) = \begin{cases} g\left(x_{w_i^{\vartriangle}}\right) & \text{if } \theta_{w_i^{\triangledown}} = \theta_{w_i^{\vartriangle}+1} \\ h\left(x_{w_i^{\triangledown}}\right) & \text{if } \theta_{w_i^{\triangledown}-1} = \theta_{w_i^{\vartriangle}} \\ \min\left\{g\left(x_{w_i^{\vartriangle}}\right), h\left(x_{w_i^{\triangledown}}\right)\right\} & \text{otherwise,} \end{cases}$$

where

$$g\left(x_i\right) = \begin{cases} f\left(x_i\right) & \text{if } \theta_{i-1} < \theta_i \text{ (define } \theta_0 = 0) \\ \min\left\{g\left(x_{i-1}\right), f\left(x_i\right)\right\} & \text{otherwise; and} \end{cases}$$

$$h\left(x_i\right) = \begin{cases} f\left(x_i\right) & \text{if } \theta_i < \theta_{i+1} \text{ (define } \theta_{n+1} = m+1) \\ \min\left\{f\left(x_i\right), h\left(x_{i+1}\right)\right\} & \text{otherwise.} \end{cases}$$

In effect, the CLSA creates m blocks using the θ_i relating to the corresponding x_i:

$$\begin{aligned}
\theta_1, \theta_2, \ldots, \theta_{b_1} &= 1 \quad \text{where } x_1, x_2, \ldots, x_{b_1} \in [x_1, x_1 + k) \\
\theta_{b_1+1}, \theta_{b_1+2}, \ldots, \theta_{b_2} &= 2 \quad \text{where } x_{b_1+1}, x_{b_1+2}, \ldots, x_{b_2} \in [x_1 + k, x_1 + 2k) \\
&\vdots \\
\theta_{b_{m-1}+1}, \theta_{b_{m-1}+2}, \ldots, \theta_{b_m} &= m \quad \text{where } x_{b_{m-1}+1}, x_{b_{m-1}+2}, \ldots, x_{b_m} \in [x_1 + (m-1)k, x_n] \,.
\end{aligned}$$

When the algorithm considers each point x_i for the minimum f in the window spanning k_0 either side, it checks whether the most extreme x-values in this window are either in the current block or one block away (these values cannot be further than one block away as block sizes are of length k) to decide on which combination of g and h are required. The recursively defined g and h vectors are similar to those used in the LSA.

Note that the occurrence of $\theta_i \neq j$ for any $i = 2, 3, \ldots, n-1; j = 2, 3, \ldots, m-1$ (empty blocks) or $x_{b_{j-1}+1} = x_{b_j}$ for any $j = 2, 3, \ldots, m$ (blocks with only one x_i) will not effect the proposed algorithm. Additionally, the CLSA is a generalisation of the LSA and can be used in its place.

Proposition 2.1. *Consider a set $X = \{x_1, x_2, \ldots, x_n\}$, $x_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$ and $x_i < x_j \ \forall i < j$, and the function f defined $\forall x_i \in X$. For a flat one-dimensional SE of size k, ϵ_B from Definition 2.10 and r_{CLSA} from the CLSA, the following equality holds $\forall x_i \in X$,*

$$r_{CLSA}\left(f(x_i)\right) = \epsilon_B\left(f\right)\left(x_i\right).$$

A proof of Proposition 2.1 can be found in Appendix B. An R implementation of the CLSA is provided in Appendix A.1.

Examples of the CLSA on continuous data

To illustrate how the CLSA works, consider two cases of the algorithm in returning the erosion in Figures 2.9 and 2.10.

Figure 2.9 shows a case where $\epsilon_B(f)(x_{12}) = 3$ using a SE of size $k = 3$. It can be seen that,

$$\theta_{w_{12}^{\triangledown}-1} = \theta_{10} = 3 \neq \theta_{w_{12}^{\vartriangle}} = \theta_{13} = 4,$$

but,

$$\theta_{w_{12}^{\triangledown}} = \theta_{11} = 3 = \theta_{w_{12}^{\vartriangle}+1} = \theta_{14}.$$

Therfore the desired result is also achieved using the CLSA as,

$$r_{\text{CLSA}}\left(f\left(x_{12}\right)\right) = g(x_{w_{12}^{\vartriangle}}) = g(x_{13}) = 3 = \epsilon_B(f)(x_{12}).$$

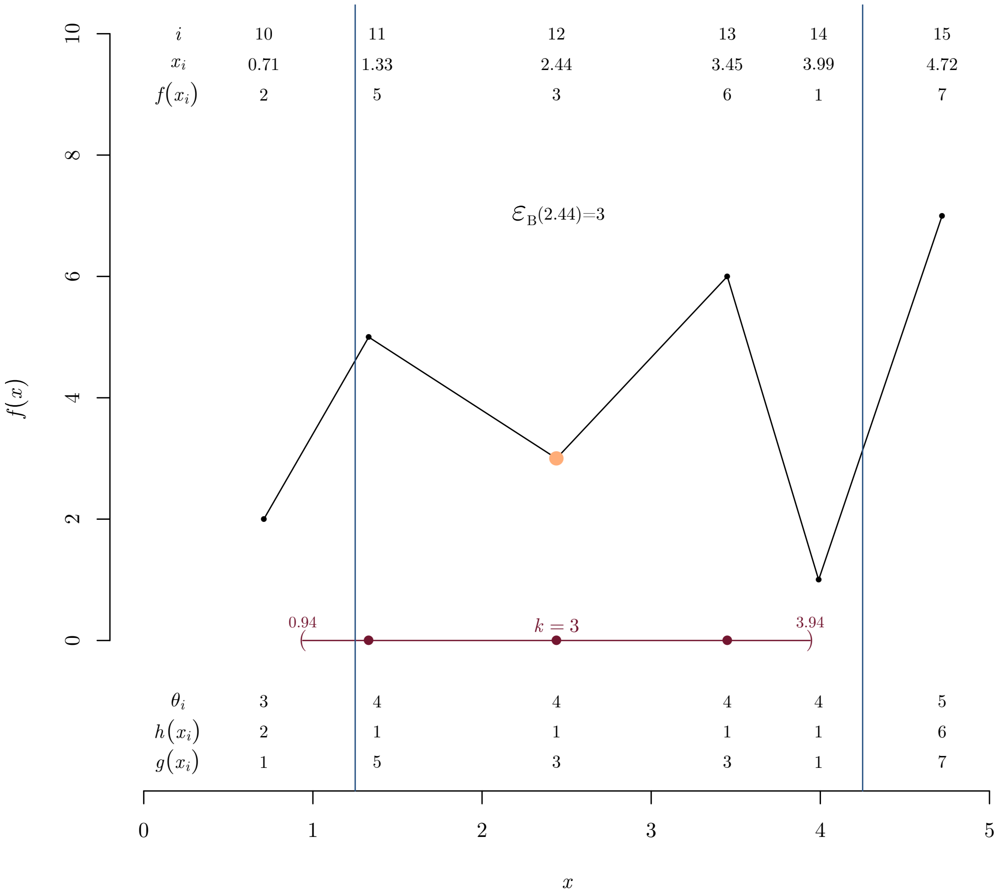


Figure 2.9: An example of data for $x_i = x_{12} = 2.44$ and $k = 3$ where $\theta_{w_i^{\triangledown}} = \theta_{w_i^{\vartriangle}+1}$ (i.e. $\theta_{w_{12}^{\triangledown}} = \theta_{w_{12}^{\vartriangle}+1} = 4$) and the computation required to return the result of the CLSA (the tan coloured point $f(2.44) = 3$).

Figure 2.10 is the different case where $\theta_{w_i^{\triangledown}-1} = \theta_{w_i^{\vartriangle}}$, as opposed to the case shown in Figure 2.9 where $\theta_{w_i^{\triangledown}} = \theta_{w_i^{\vartriangle}+1}$. To obtain the erosion of point $x_i = x_9 = 2.44$ for $k = 3$ using the CLSA, observe that

$$\theta_{w_9^{\triangledown}} = \theta_8 = 3 \neq \theta_{w_9^{\vartriangle}+1} = \theta_{11} = 4,$$

and

$$\theta_{w_9^{\triangledown}-1} = \theta_7 = 3 = \theta_{w_9^{\vartriangle}} = \theta_{10}.$$

Therefore, the result of the CLSA erosion is

$$r_{\text{CLSA}}\left(f\left(x_9\right)\right) = h(x_{w_9^{\triangledown}}) = h(x_8) = 3.$$

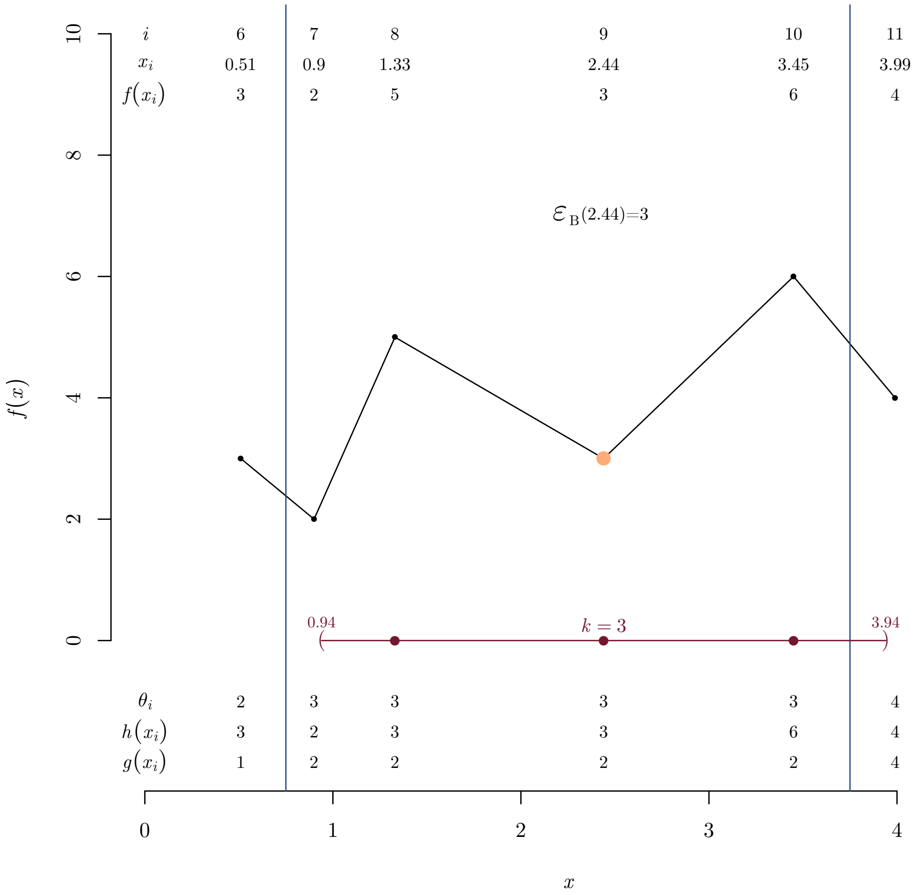


Figure 2.10: An example of data where $\theta_{w_i^{\triangledown}-1} = \theta_{w_i^{\vartriangle}}$ and the computation required for the continuous line segment algorithm.

Efficiency of algorithm

The naive algorithm to find moving window minimum consists of the linear-time process of finding the indexes of points at the upper and lower edges of the sliding window for each element, by incrementing the edge indexes from the previous element when required. Using a_k as the average number of data points in the sliding window of size k, the computational cost of finding the minimum value in the window requires approximately $a_k - 1$ comparisons per element. This is because each element requires, on average, a minimum comparison of all the data points in the window except one: the first datapoint does not require a comparison. The resulting computational complexity is $\mathcal{O}\left(a_k n\right)$ for naive algorithm; dependent on the size of the sliding window and the number of elements in X.

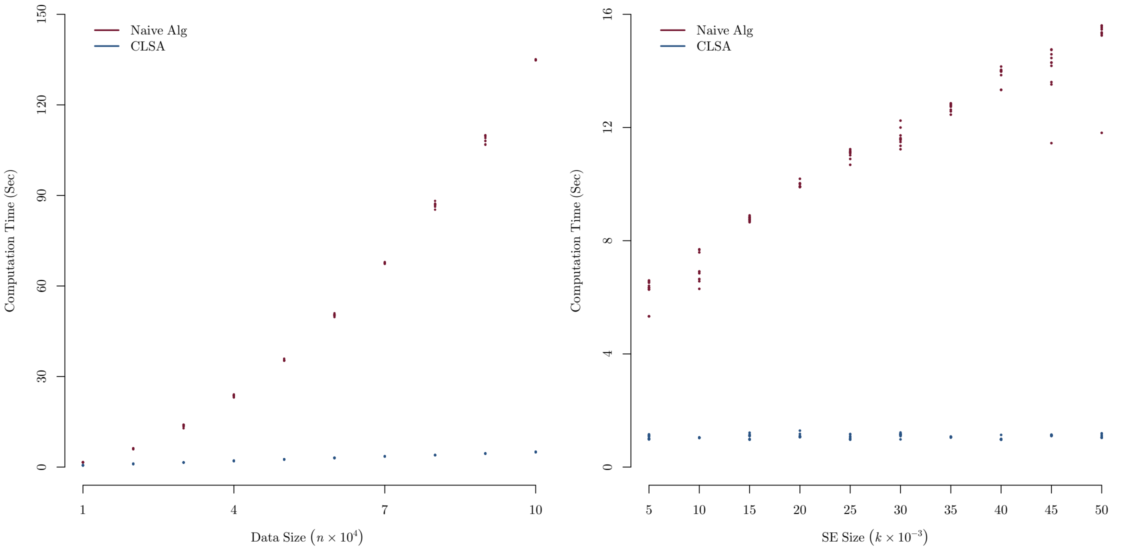
Like the LSA, the CLSA is a linear-time algorithm irrespective of the window size, k. For the CLSA, a linear-time progression through the n elements is required to assign integers of the Θ vector, as each element is an integer equal to or greater than the integer that precedes it. The linear-time process of finding the $W^{\triangledown}$ and $W^{\vartriangle}$ indexes at the lower and upper edges of the sliding window, respectively, for each element is required similar to the naive algorithm. One linear-time sweep forward and back of the data is required to create g and h each. A final sweep of the created vectors $W^{\triangledown}$, $W^{\vartriangle}$, Θ, g and h is required to compute the r_{CLSA} values. Each $r_{\text{CLSA}}\left(f\left(x_i\right)\right)$ calculation requires the tests $\theta_{w_i^{\triangledown}} = \theta_{w_i^{\vartriangle}+1}$, $\theta_{w_i^{\triangledown}-1} = \theta_{w_i^{\vartriangle}}$ or $\min\left\{g\left(x_i\right), h\left(x_i\right)\right\}$. It can therefore be deduced the CLSA is $\mathcal{O}(n)$ complexity requiring a series of linear-time operations, importantly independent of the length of the sliding window, k.

Given the MS application, $a_k - 1$ operations per element in the naive algorithm would be much larger than the constant number of operations required per element for the CLSA and efficiency strongly favours the CLSA. It should be pointed out the CLSA requires extra memory availability beyond the naive iterative algorithm for the creation of the vectors $W^{\triangledown}$, $W^{\vartriangle}$, Θ, g and h.

Another computational advantage of the CLSA is by using the minimum of the two temporary vectors g and h as opposed to the minimum of a non-constant number of data points for each $x_i \in X$, vectorised programming can be utilised instead of loops. This is of significant advantage in programming languages that are interpreted like R.

Figures 2.11(a) and 2.11(b) show times required to compute morphological openings (erosion and a subsequent dilation) using the naive algorithm and the CLSA for randomly generated data.[6] The data consisted of x values randomly generated from a uniform$(0, 1)$ distribution (sorted in ascending order) and positive f intensity values from a uniform$(0, 20)$ distribution. Figure 2.11(a) shows the computational time to calculate a morphological opening for data generated with constant SE size $(k = 0.1)$. Figure 2.11(b) shows the computational time for calculating the morphological opening for generated data $(n = 20000)$ with varying SE sizes. Importantly, these figures demonstrate the CLSA is a linear-time algorithm with respect to the data size. Figure 2.11(a) shows the CLSA is far superior to the naive algorithm when the number of points contained in the moving window is large. i.e., when a_k increases, so does the computational time of the naive algorithm but not the CLSA.

[6] Using a MacBook Pro7 (Intel Core 2 Duo 2.4 GHz, 3 MB L2 Cache, 8 GB Memory).



(a) Varying data size ($k = 0.1$)

(b) Varying SE size ($n = 20000$)

Figure 2.11: Computation time of calculating the morphological opening for randomly generated data in R.

2.2.7 Comparison of piecewise and continuous baseline subtraction

From the previous sections, if a transformation of MS data to create roughly uniform peak widths is known, an efficient and effective baseline correction can be performed using the CLSA. Figures 2.12, 2.13 and 2.14 show a randomly selected spectrum for the GC mice data and the comparative baseline estimates using the standard piecewise method of baseline subtraction (morphological opening) and the transformation of the TOF-axis with the use of the novel CLSA. Using the transformed data of x^{*}=TOF$^{1/4}$, the top-hat operator has the slight tendency to be more conservative at lower m/z-values and possibly undercut the peaks at higher m/z-values than the standard piecewise treatment. The piecewise approach is performed manually (by inspection) so the transformed and continuous approach may suffer from a reduction in sensitivity in comparison. The trade-off between exactness of the piecewise approach to the speed of the automated transform and continuous approach may be a consideration, especially if a known relationship exists between the peak width and peak location. As stated previously, some literature (Siuzdak, 2006; House et al., 2011) has suggested the relationship between peak width and peak location to be roughly linear, but this was not observed for these data.

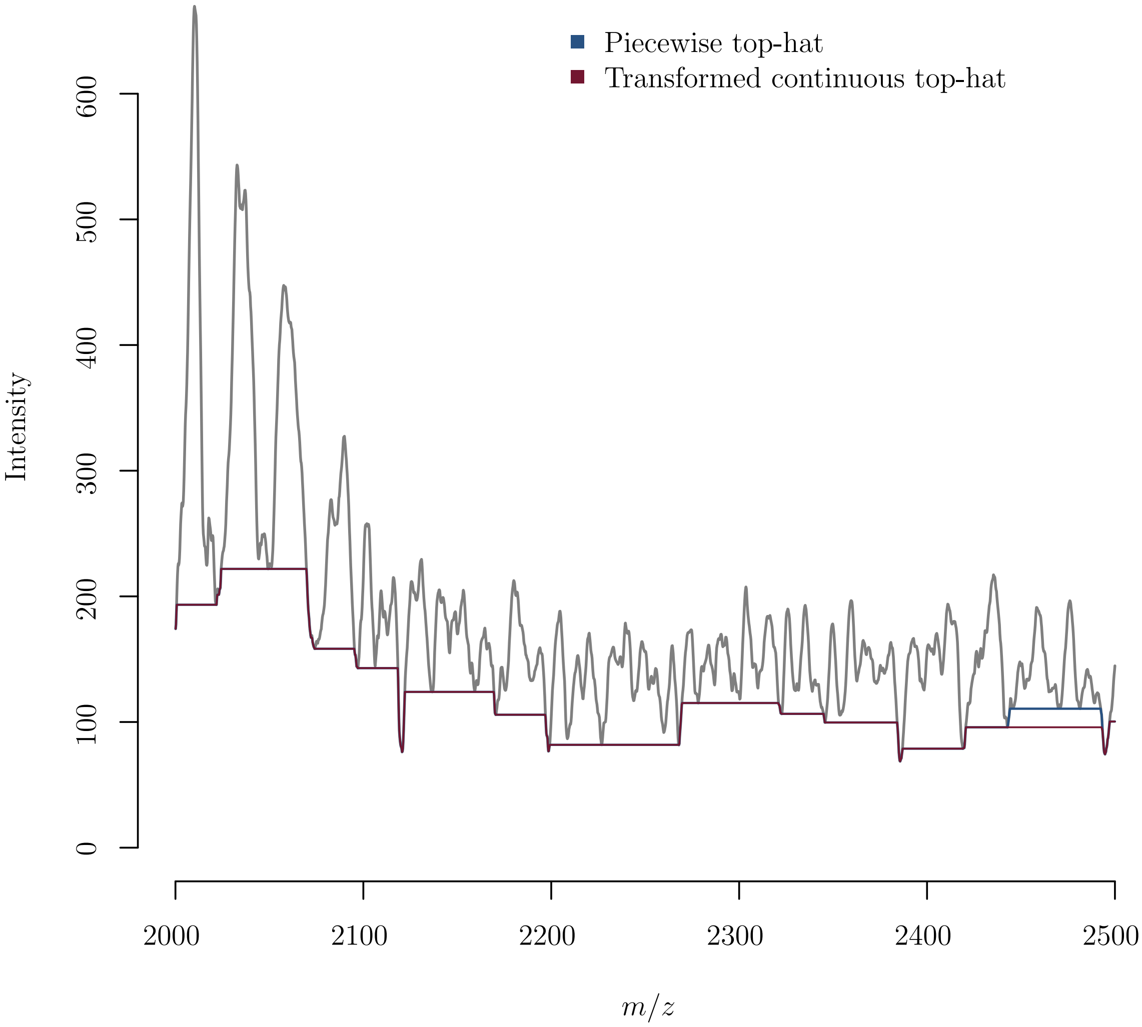


Figure 2.12: Baseline correction on a randomly selected spectrum at the low end of m/z-values from the GC mice dataset using piecewise top-hat and transformed TOF-values with constant SE width and the continuous definition of morphological operators.

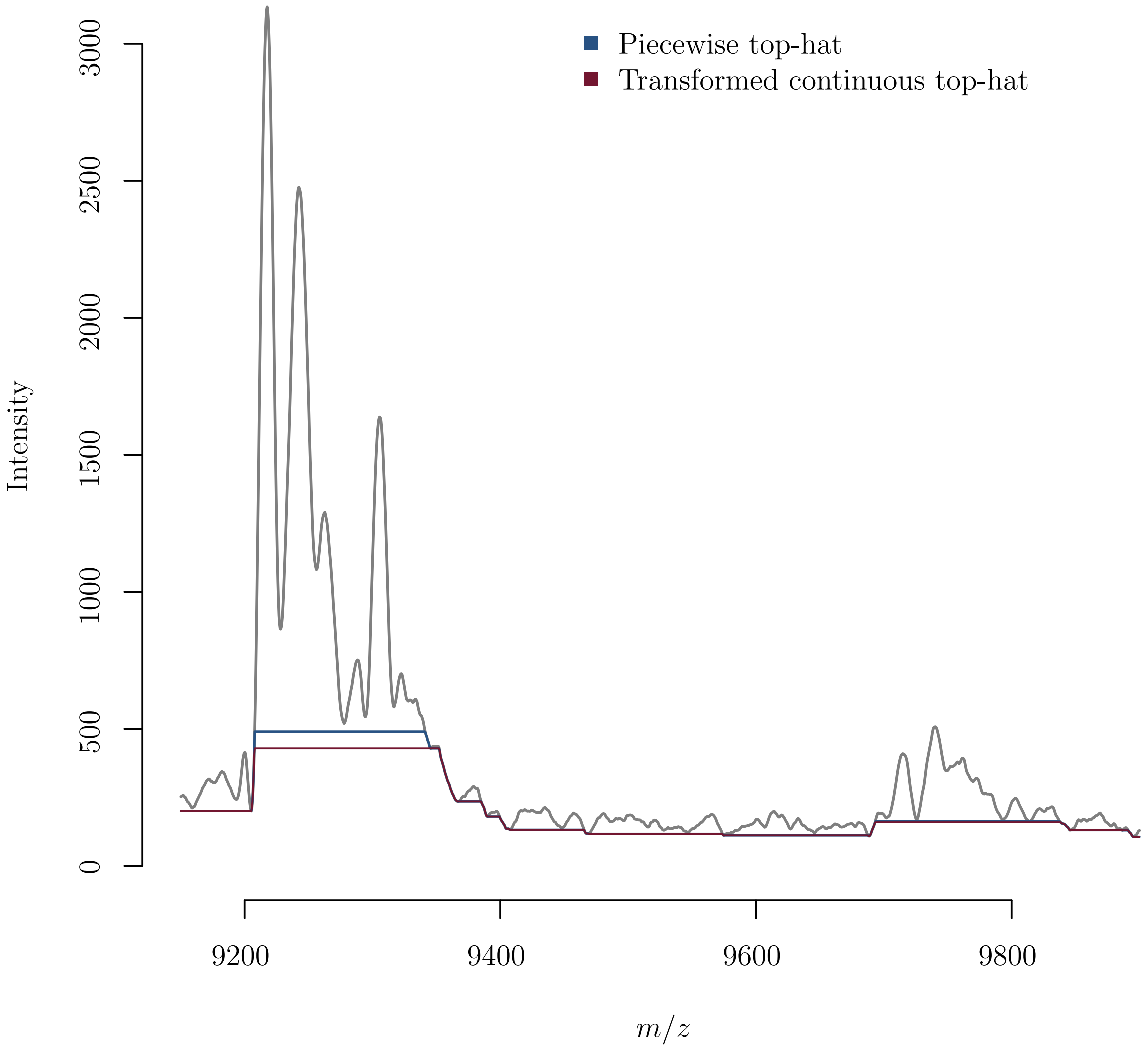


Figure 2.13: Baseline correction on a randomly selected spectrum at an intermediary section of m/z-values using piecewise top-hat and transformed TOF-values with constant SE width and the continuous definition of morphological operators.

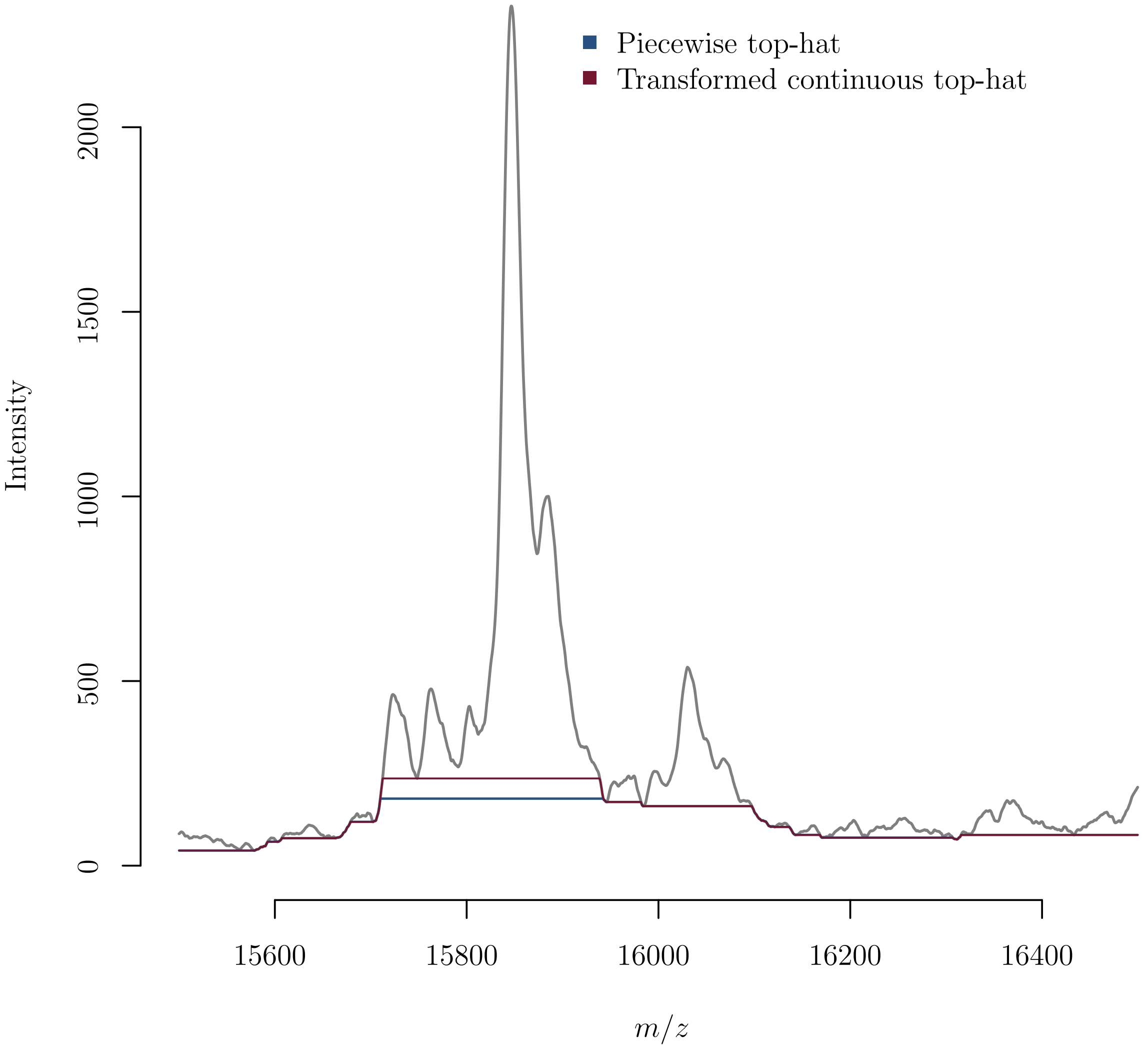


Figure 2.14: Baseline correction on a randomly selected spectrum at the high end of m/z-values using piecewise top-hat and transformed TOF-values with constant SE width and the continuous definition of morphological operators.

2.3 Recommendations

Presented here are methods of proteomic TOF-MS signal adjustment for the purpose of creating data which ideally contain signals from biological molecules only. Every step in pre-processing is important and less precise methods in early pre-processing may have amplified unwanted effects at the analysis stage. There is currently no standard pipeline for MALDI/SELDI TOF-MS data pre-processing.

Savitzky-Golay smoothing yielded superior results to standard smoothing. Although not a new method, its use is advocated here as a pre-processing smoothing algorithm to remove electrical noise in mass spectra without diminishing existing peaks, which is very important as signals of interest may not be the most pronounced of the raw spectra.

Morphological openings and the top-hat operator are an ideal choice for baseline subtraction and has been applied effectively here on MS data. The top-hat operator is supported by limited software packages but its properties make it an ideal method. A largely overlooked algorithm to speed up top-hat computation is also given, increasing the desirability of the method. Additionally and most importantly, a novel method of automating much of the baseline subtraction process is presented with a new algorithm for its computation.

Chapter 3

Methods of inter-spectra pre-processing

Following intra-spectra pre-processing, spectra must be normalised and peak-aligned to allow meaningful comparisons of proteomic profiles in the downstream analysis. Normalisation is the process of adjusting the arbitrary intensities present in each raw spectrum to a common scale. Each spectrum is then considered to represent an equal amount of ionised analyte from each MALDI/SELDI spot and reaching the detector, despite this not being achievable physically by the TOF system. Here, available spectra normalisation methods are compared to empirical quantile and cyclic LOESS normalisation which are used in microarray pre-processing but not in protein analysis. Peak alignment, combined with peak detection, is the final pre-processing step to extract protein mass and expression information. A new method of alignment is proposed, modifying current dynamic programming algorithms. This dynamic programming method is then compared to a standard peak alignment technique.

3.1 Pre-processing step III: spectra normalisation

Up to this point, the pre-processing steps on the data have been intra-spectra in the sense that no adjustments have been made to correct batch-effects from spectrum to spectrum. Normalisation is the first of the inter-spectra adjustments required.

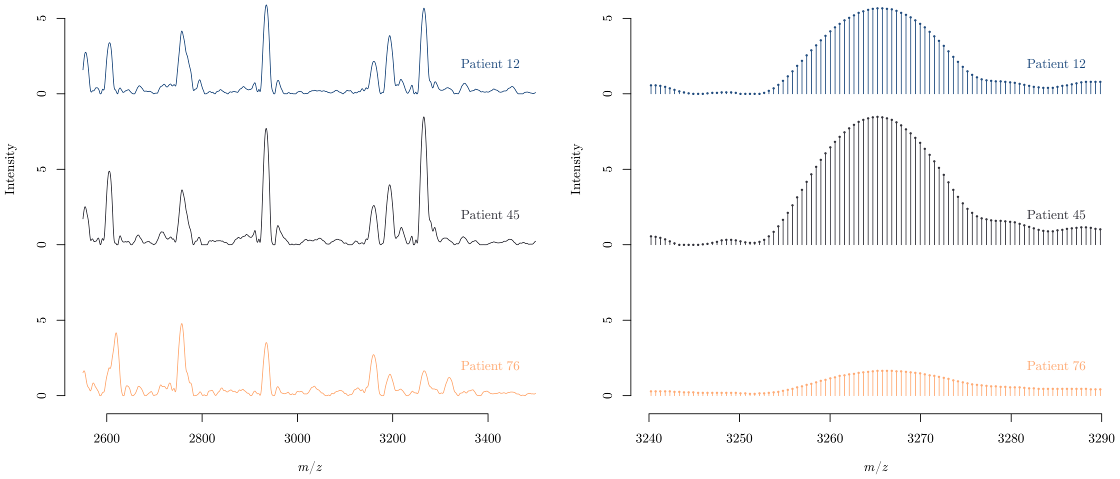
Proteomic MALDI/SELDI TOF-MS suffers from variability produced by the MS technology itself in addition to sampling variability. Normalisation is a method to help ensure the spectra signal in mass spectra are reflective of true peptide expression. Normalisation aims to adjust the signals so that peptides across spectra are appropriately proportionate.

Many simple methods of spectra normalisation have been proposed, including the mapping of intensity percentiles within spectra to $[0, 1]$ (Randolph, 2006) or the transformation of a spectrum's intensities via subtraction of the mean and division by the standard deviation (Randolph, 2006; Meuleman et al., 2008; Gong et al., 2012). However, simple normalisation via total ion current (TIC) has emerged as the standard normalisation method for TOF-MS (Fung and Enderwick, 2002; Sauve and Speed, 2004; Ressom et al., 2005; Gong et al., 2012).

Normalisation using TIC is based on the assumption that the proteomic profiles of samples should contain close to the same amount of total peptide per sample. i.e., any observed difference in the total amount of received signal in each spectrum is an artefact of the system, in which desorbed ions are not necessarily reflective of the peptide content of the sample itself. Theoretically, differences in sample profiles should only be up- and down-regulated peptides. The up- and down-regulated peptides in a spectrum, with respect to any given profile, should therefore be equal in aggregate and thus their effect on the total of peptide expression in the spectrum is nullified. For these reasons, TIC normalisation (TCN) has good overall normalisation characteristics but will not satisfactorily adjust individual peaks or intervals within individual spectra. Thus TCN will be referred to as a global normalisation method.

Two alternative methods have been investigated in this research; firstly, empirical quantile normalisation (EQN) and secondly, cyclic LOESS normalisation (CLN). Importantly, these two methods offer an approach to normalisation which can account for systematic differences in peptide expression in local areas of the spectrum and will be referred to as local normalisation methods.

When EQN and CLN have been previously applied to microarray data, the two methods have yielded similar results (Bolstad et al., 2003), or provided results slightly in favour of EQN (Ballman et al., 2004). Computational efficiency is an additional consideration. EQN requires considerably less total computational time than compared to CLN. Even though proteomic mass spectrometry and genetic microarray experiments produce high throughput biological data, there are also important differences. In microarray experiments, normalisation is applied to the probe signals which are the features of interest in the statistical analysis. For proteomic mass spectra, the complete set of discretely measured intensities across the entire m/z-axis are not the features intended to be analysed statistically. Peak expressions within spectra are the features of interest and are a subset of the intensities measured across the m/z values, a set orders of magnitude smaller in size. Peak expressions are not obtained until later in the pre-processing. Proteomic mass spectra data also have a different underlying structure. Successive intensity values along the m/z-axis in proteomic spectra (see Figure 3.1 for example) are highly correlated because they are discrete measurements of a theoretically continuous underlying signal. This same structure is not necessarily present in microarray data. Finally, CLN requires additional consideration for mass spectra because of the high abundance of zero intensities. CLN relies on taking logarithms of intensities for which zero values are undefined. Thus, structural differences between genetic microarray data and proteomic mass spectra data imply the results of Bolstad et al. (2003) and Ballman et al. (2004) are not guaranteed to hold for proteomic mass spectra data.



(a) A subset of the m/z-axis **(b)** A zoomed view of Figure 3.1(a)

Figure 3.1: Randomly selected spectra from the Adam et al. (2002) dataset.

Spatial and run-order biases play a role in observed MS signal. Use of run-order and spatial variables have largely been ignored in the evaluation of effective MS normalisation. Graphical comparisons using MALDI chip location will be used to evaluate the effectiveness of the global and local normalisation techniques to make suitable adjustments on subsections of the spectra.

3.1.1 Total ion current normalisation

TCN is performed by summing the total intensities (discrete expressions as seen in Figure 3.1) for each spectrum. Then a reference spectrum or mean spectrum TIC is calculated. The intensity vector of each spectrum $\boldsymbol{F}_i,\ i = 1, 2, \ldots, n$ is adjusted by multiplication of the reference or mean TIC and divided by its own TIC. This process can be represented by the following reassignment of the intensities,

$$\boldsymbol{F}_i^* \leftarrow \frac{\frac{1}{n}\sum_{j=1}^{n} C(\boldsymbol{F}_j)}{C(\boldsymbol{F}_i)}\boldsymbol{F}_i \qquad \forall i = 1, 2, \ldots, n,$$

where $C(\boldsymbol{F}_i) = \sum_{t=1}^{T} F_{it}$ is the total count of the intensities in spectrum $\boldsymbol{F}_i = [F_{i1}\ \ F_{i2}\ \ldots\ F_{iT}]$.

Theoretically, differences between sample profiles should be attributable solely to up- and down-regulated peptides. TCN assumes that these differences do not make a significant contribution to the TIC.

3.1.2 Empirical quantile normalisation

Here, we propose the use of EQN on MS data. This normalisation method has been applied successfully in the microarray literature (Bolstad et al., 2003). The motivation for this normalisation method is not only that the total TIC should be roughly equal for each of the spectra but that the intensities contained within each spectrum should be derived from the same distribution. EQN is performed by the following process:

(1) Order the intensity values in each spectrum from the smallest to largest intensity.

(2) Replace the ordered intensity values with the mean or median intensity at that ordered position (position $1, 2, \ldots, T$) calculated across all the spectra.

(3) Re-order the individual spectra intensities back to their original positions returning them back to their corresponding m/z-values.

An illustration of the EQN process can be seen in Figure 3.2. Upon EQN, the spectra have different intensities at different m/z-values but the distribution of intensities is exactly the same for each spectrum. An R-implementation is provided in Appendix A.2.

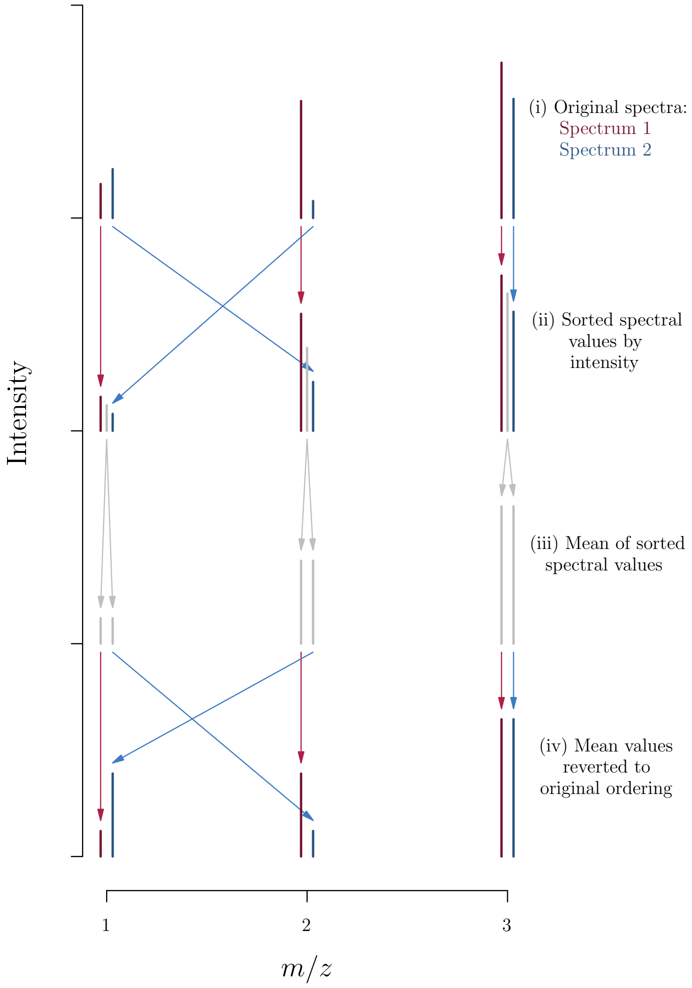

(i) Original spectra:
Spectrum 1
Spectrum 2
(ii) Sorted spectral values by intensity
(iii) Mean of sorted spectral values
(iv) Mean values reverted to original ordering
Intensity
1
2
3
m/z

Figure 3.2: An example empirical quantile normalisation on two spectra with intensities at three m/z-values.

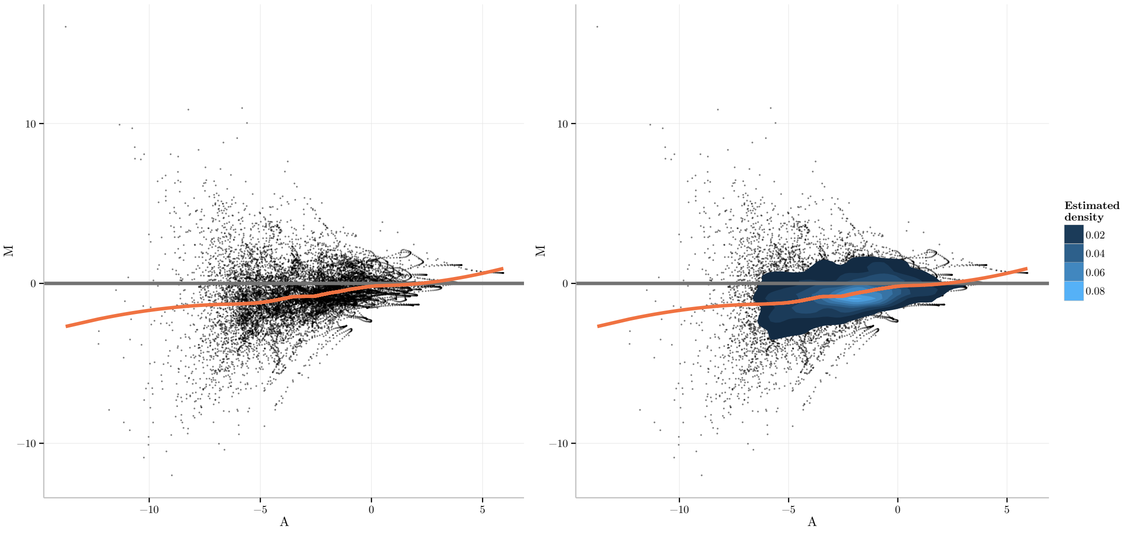
3.1.3 Cyclic LOESS normalisation

CLN offers an alternative for normalisation that can also account for systematic differences in peptide expression across spectra using local adjustments. CLN is used in the microarray literature (Yang et al., 2002; Quackenbush, 2002; Smyth and Speed, 2003; Bolstad et al., 2003). Sauve and Speed (2004) considered a transformation of the MS data and observed the potential for correcting systematic bias using LOESS. Presented here is CLN, an extension of the transformation and LOESS method used in Sauve and Speed (2004), which is novel to MS normalisation.

Local adjustment for peak intensities that may be artefacts of the system (such as desorption or detection) is performed by first transforming the data. Consider a pairwise comparison of two spectra. Let the first spectrum, i, have intensity F_{it} at mass t and similarly F_{jt} for the second spectrum. Now define the quantities

$$\begin{aligned} M_t &= \log_2 F_{it} - \log_2 F_{jt}, & (3.1)\\ A_t &= \tfrac{1}{2}\left(\log_2 F_{it} + \log_2 F_{jt}\right). & (3.2) \end{aligned}$$

M_t is the alias for minus the log intensity differences but is best thought of as the log ratio of intensities across spectra for a particular mass. A_t is an alias for the average log intensity across spectra for a particular mass. In effect, a transformation of the data to a different scale is used to assess the relative differences in expression based on mass location. For the moment F_{it} values are assumed non-zero, the effect of zero F_{it} values is discussed later.



(a) Scatter plot

(b) Scatter plot with density overlay

Figure 3.3: Identical MA plots of two spectra from the Adam et al. (2002) data. The density plot better illustrates the distribution of points.

If (A_t, M_t) pairs are calculated $\forall\, t = 1, 2, \ldots, T$, a plot of M_t vs. A_t will provide an indication of whether a mass-location bias may exist when comparing the intensities of two spectra; this is the 'MA' plot seen in Figure 3.3. It is of course not expected that intensities will be the same for both spectra (i.e. log ratio of 0) at the same mass, or at every A_t location, but the scatter should be distributed around zero. If the scatter is not centred around zero, this suggests there is some location or mass dependency in intensities. A LOESS line can be fitted through the data to check the location of the average scatter across all A_t.

The M_t values can be adjusted to M_t^* by simply assigning

$$M_t^* \;=\; M_t - \ell(A_t) \qquad (3.3)$$

where ℓ is the LOESS regression function. Note the A_t values are not changed.

To obtain the adjusted F^*_{it} and F^*_{jt} values, the adjustments on the M-scale need to be transformed back to the original scale. Re-arranging (3.1),

$$\log_2 F^*_{it} = M^*_t + \log_2 F^*_{jt} \tag{3.4}$$

and (3.2),

$$\log_2 F^*_{it} = 2A_t - \log_2 F^*_{jt}, \tag{3.5}$$

to then equate (3.4) and (3.5),

$$\begin{aligned} M^*_t + \log_2 F^*_{jt} &= 2A_t - \log_2 F^*_{jt} \\ \Leftrightarrow F^*_{jt} &= 2^{A_t - M^*_t/2}. \end{aligned} \tag{3.6}$$

To obtain the adjusted F_{it}, F^*_{it} values, substitute (3.6) into (3.4), then

$$F^*_{it} = 2^{A_t + M^*_t/2}. \tag{3.7}$$

From this, the intensity differences are corrected in this transformed space then transformed back using Equations (3.6) and (3.7).

MS data have many intensity points of zero for which the logarithm will not be defined. Zero values will be omitted from normalisation since it is not desirable to adjust non-existent signals . In the case where both F_{it} and F_{jt} equal zero, no adjustments or transformation will be made, as desired. If only one of F_{it} or F_{jt} equals zero, then the other value will not be adjusted. This might seem like an issue as a non-zero F_{it} or F_{jt} is not adjusted, but a pair of intensities is required to make a comparison in the transformed MA-space to make the adjustment meaningful. The 'orphaned' F_{it} or F_{jt} non-zero value will be adjusted against other non-zero intensities when other pairwise combinations of spectra are considered. Please refer to Appendix A.2 for an R-implementation of a pairwise-MA adjustment of two vectors of intensities.

Figure 3.4 illustrates the effect of MA-LOESS adjustment for only two spectra. When the spectra are transformed to the MA-space, LOESS adjusted and back-transformed, peaks are adjusted and m/z-locations of no or little signal are not adjusted.

How to perform MA-LOESS normalisation for more than two spectra needs to be considered. To handle the $n > 2$ case, pairwise comparisons can be performed but require all possible pairwise combinations to be considered. This requires $\binom{n}{2} = \frac{n(n-1)}{2}$ comparisons. The $n-1$ adjustments on the transformed (MA) scale for each spectrum are averaged and back transformed using an average M^*_t for Equations (3.6) and (3.7). This process is iterated until a threshold of minimum change is achieved, thus the naming, cyclic LOESS normalisation. Such a method is quite computationally intensive and may be prohibitive for a large number of spectra.

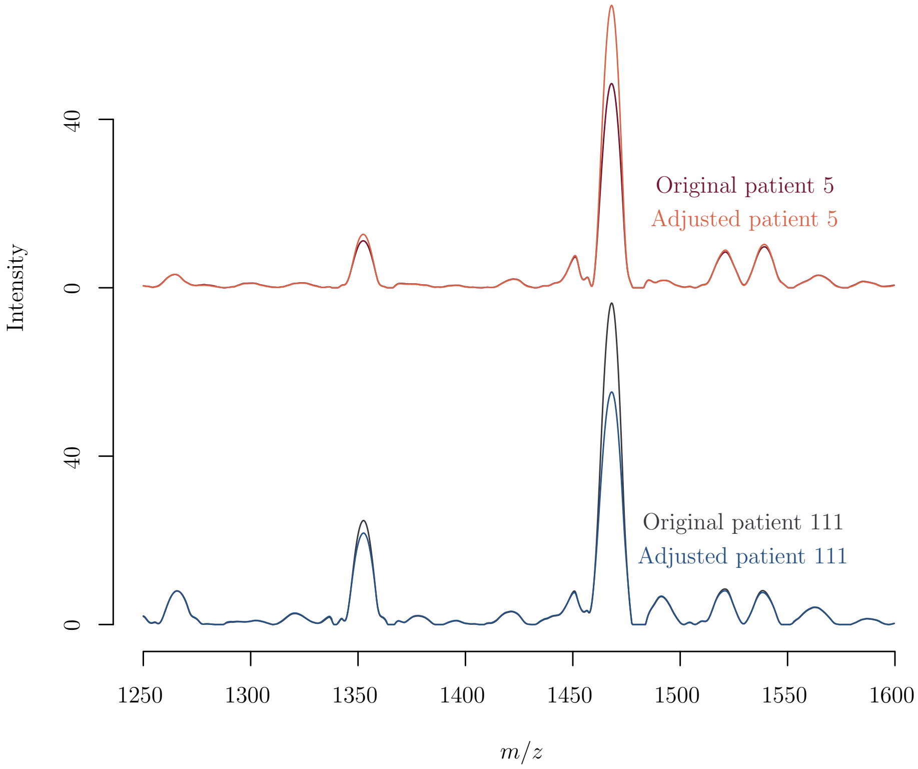


Figure 3.4: The effect of MA-LOESS adjustment on two spectra.

3.1.4 Evaluating methods of normalisation

Meuleman et al. (2008) suggest that spectra normalisation is the optimisation of two simultaneous objectives. The first being the minimisation of inter-spectra variance and the second being the increase in differentiation between experimental groups. They suggest the first objective can be assessed by the (minimisation of) coefficient of variation (CV) and the second by the (maximisation of) correct classification of models predicting the experimental group, under the assumption of equal cost of correct and incorrect classification.

In the first instance, these two criteria will be used to evaluate the performance of the three methods of normalisation TCN, EQN and CLN. Another approach to investigate whether there has been a reduction in spectra noise is to observe the TIC of spectra along subsets of the m/z-axis. As the assumption of TCN is that up- and down-regulated proteins should roughly cancel each other out, this assumption can be tested on smaller subsets of the m/z-axis as long as they are sufficiently wide. For example, if the m/z-axis is divided roughly into three sections, the TIC for each section across all spectra should be roughly the same after normalisation. Figure 3.5 shows the TIC of asthma2 dataset spectra for the subset of the m/z-axis 4000-7000Da (roughly the middle third of m/z-values) and the reduction in noise associated with the three different normalisation methods. It can be seen that both EQN and CLN reduce the variability in TIC over the axis subset far more than TCN or no normalisation.

Coefficient of variation

A robust version of the traditional coefficient of variation, $CV = \sum_{i=1}^{n} s_i/m_i$, has been proposed by Meuleman et al. (2008),

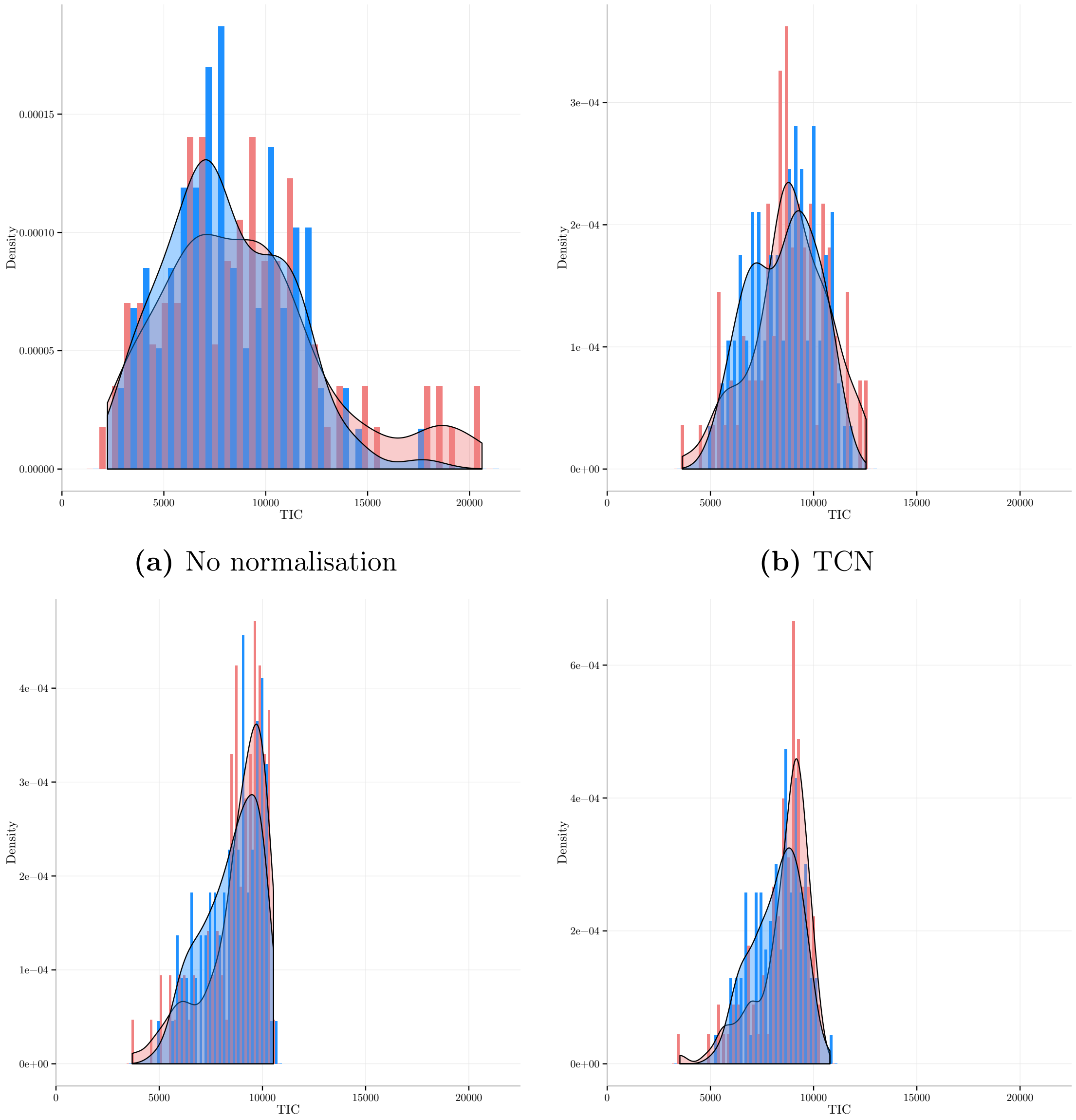
$$CV_r = \frac{\sum_{i=1}^{n} s_i m_i}{\sum_{i=1}^{n} m_i^2},$$

where m_i is the mean of the peaks and s_i is the standard deviation of the peaks in spectrum i. The rationale is that the traditional CV is susceptible to instability for small values of m_i, as may be the case for TOF-MS. Just like the traditional CV, smaller CV_r is preferred and suggests a lower level of variability. The CV_r can be calculated on 'spiked in' peaks in spectra, but unfortunately these peaks are not present in any of these data. The CV_r is thus estimated using detected peaks where the method for finding these is outlined in §3.2.

CV_r is a scale-free statistic for estimating the variability of peak intensity after normalisation. Table 3.1 shows EQN generally reduces this variability compared to standard TCN. The asthma2 dataset is an exception, which as discussed in §1.3.5, is particularly noisy with sample degradation. The de Noo et al. (2006) dataset, as discussed in §1.3.4, had already been pre-processed so the additional normalisation had little to no effect on peak variability. It is unknown what prior normalisation method was used. The CV_r was improved for the remaining datasets when EQN was used when compared to TCN. On the datasets that underwent normalisation using CLN, the CV_r when using EQN was superior to the CV_r when using CLN. It became apparent that CLN was less effective than EQN and given the additional computational complexity (days of computation opposed to seconds using naive `R` code) it was not pursued further.

Classification signal

The methods used to determine the classification signal will be explained in Chapters 5 and 6. However, to provide an indication of performance in classification, as recommended by Meuleman et al. (2008), using the two methods of normalisation (TCN and EQN), Table 3.2 gives the mean misclassification proportion for the test data used in each case (independently held data from the model creation data §5.1). Percentile bootstrap 95% confidence intervals (Davison and Hinkley, 1997)



(a) No normalisation

(b) TCN

(c) EQN

(d) CLN

Figure 3.5: Density and histogram plots of total ion current for asthma2 spectra over the 4000-7000 subsection of the m/z-axis with constant x-axes. Coral and blue represent female and male spectra, respectively.

Table 3.1: CV_r value using different normalisation methods and data.

		Normalisation method		
Dataset	n	TCN	EQN	CLN
Morris et al. (2005) (synthetic)	40	2.16	1.99	2.08
Adam et al. (2002)	326	3.46	3.41	3.51
Asthma1	243	4.60	4.51	NA$^\natural$
Asthma2	195	2.02	2.09	2.10
de Noo et al. (2006)	112	4.35	4.35	NA†
GC mice	1080	7.47	7.23	NA††

†Not pursued as the de Noo et al. (2006) data was already pre-processed prior to being obtained. Further normalisation showed no marked effect.

††Computationally prohibitive for the sample size.

$^\natural$CLN was deemed to not be worthwhile pursuing from further investigation prior to evaluation of this data.

are also provided to indicate the variability in the mean misclassification. The classification errors were sampled 100,000 times with replacement and the distribution of the mean misclassification was used to obtain the percentile bootstrapped confidence intervals. Misclassification is incorrectly estimated group membership by the model. Obviously, maximised classification signal would result in minimised misclassification.

Table 3.2: Mean misclassification proportion (bootstrapped 95% confidence interval of the mean) using the normalisation methods TCN and EQN.

	Normalisation method	
Dataset	TCN	EQN
Adam et al. (2002)	0.182 (0.175, 0.189)	0.188 (0.181, 0.195)
Asthma1	0.326 (0.317, 0.336)	0.324 (0.315, 0.333)
Asthma2	0.277 (0.266, 0.287)	0.277 (0.266, 0.288)
de Noo et al. (2006)	0.054 (0.046, 0.061)	0.054 (0.046, 0.061)
GC mice	0.044 (0.038, 0.051)	0.040 (0.034, 0.046)

There is no clearly preferred method with respect to the classification signal comparing TCN and EQN (Table 3.2). It is important that the classification error is constant irrespective of the normalisation method for the de Noo et al. (2006) dataset. The normalisation methods are, for all practical purposes, redundant for the de Noo et al. (2006) data because of the prior pre-processing undertaken on it. The asthma datasets show very poor classification, indicative of the lack of inherent classification signal, and neither of the normalisation methods were able to improve it. The two remaining datasets had slightly different mean classification error when comparing the normalisation methods. The GC mice had a slightly better mean classification proportion when using EQN and the Adam et al. (2002) dataset had slightly worse mean classification proportion when using EQN. From these results it would be inferred that while EQN reduces the variability seen in the peak intensities across spectra, TCN is a sufficient normalisation method for discrimination.

Figure 3.6 utilises run-order information as to check for batch effects as well as investigate the TCN assumption that signals of up- and down-regulated protein should cancel out. Under the different normalisation methods, it is apparent that EQN and CLN remove much variability in TIC for subsections of the m/z-axis. There does seem to be a small area in the upper-left corner of each of the heatmaps with less TIC, suggesting the respective spectra need to be followed closely throughout the pre-processing and analysis steps for undue influence on results. The extreme TIC value at position (J,5) in Figure 3.6(c) for CLN would appear to be anomalous, as it was reduced by EQN to a TIC value consistent with other spectra.

M vs. A visualisations can be used as a check for experimental bias whether or not EQN is undertaken. Additionally, visualisation of batch-order effects is important to ensure biased data do not reach the discrimination stage of the analysis.

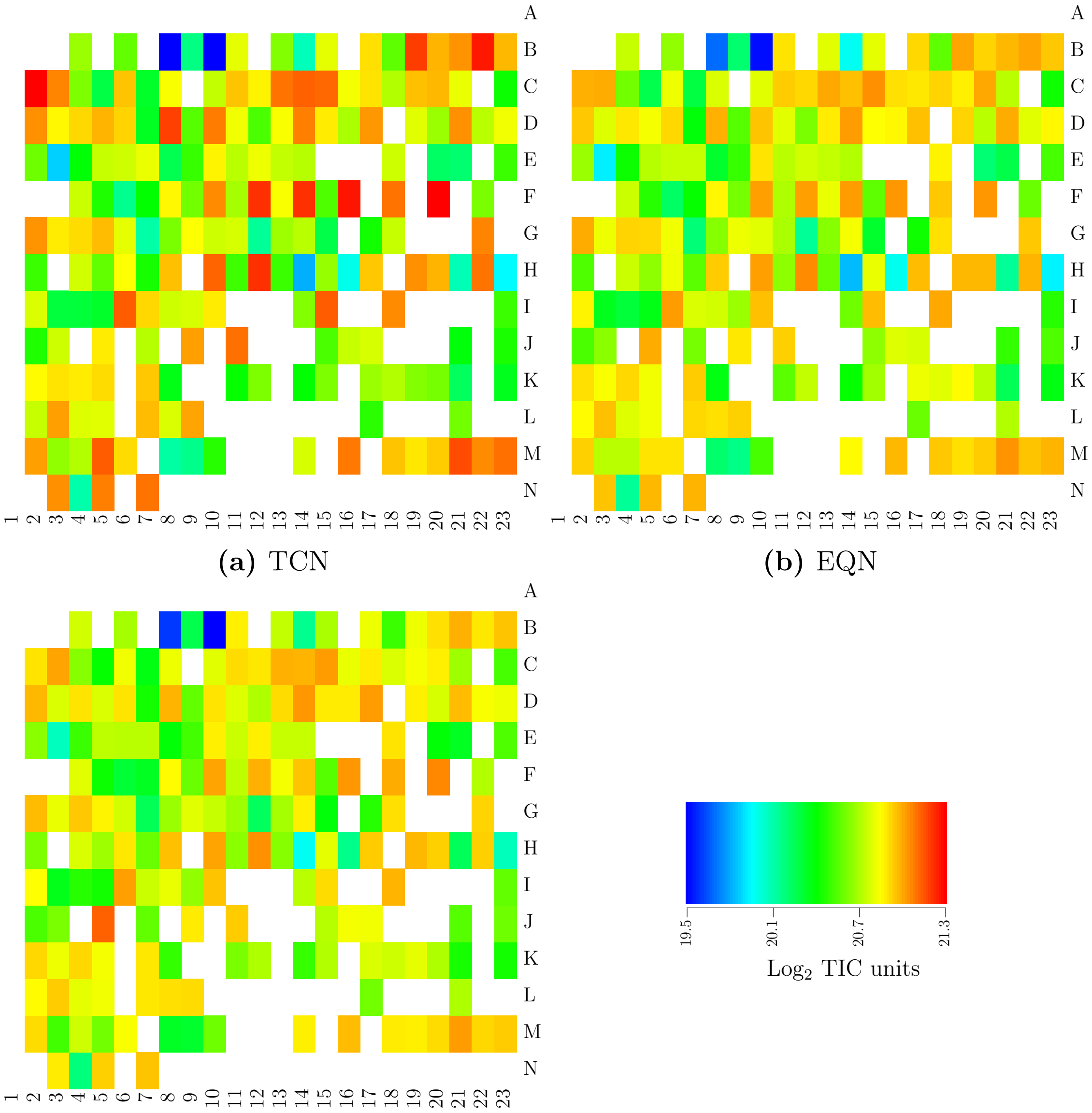
3.2 Pre-processing step IV(a): peak detection

Before peaks that are common across spectra can be identified by peak alignment, the peaks themselves need to be detected. Very simple but effective methods exist, such as the signal to noise (S2N) ratio (Tibshirani et al., 2004; Gong et al., 2012).

The S2N ratio method of peak detection involves finding spectra intensities, as a ratio to the local noise around it, exceeding a specified threshold. Intensities exceeding the threshold also must also be the maximum value in a defined small window to be considered a ‘peak’. This prevents all points on a peptide distribution being identified as peaks and limits identified peaks to a single point. An illustration of this peak-finding method is given in Figure 3.7.

The parameters of this algorithm, S2N ratio threshold and peak window width, need to be carefully calibrated and this can be performed by visual inspection. As with many of the pre-processing methods, performing piecewise peak detection over segments of the TOF- or m/z-axis is advised. Noise levels are not necessarily constant over the m/z-axis and peaks tend to be wider for increases on the TOF-axis (see Figure 3.8), thus different peak detection parameters are required for different m/z locations.

The width of peaks can be computed in a manner similar to the S2N method to detect peak vertices. By taking the negative value of the intensities, the same ‘peak



(a) TCN

(b) EQN

(c) CLN

Figure 3.6: Heat map plots of MALDI chip location total ion current for asthma2 spectra over the 4000-7000 subsection of the m/z-axis using (a) TCN, (b) EQN and (c) CLN.

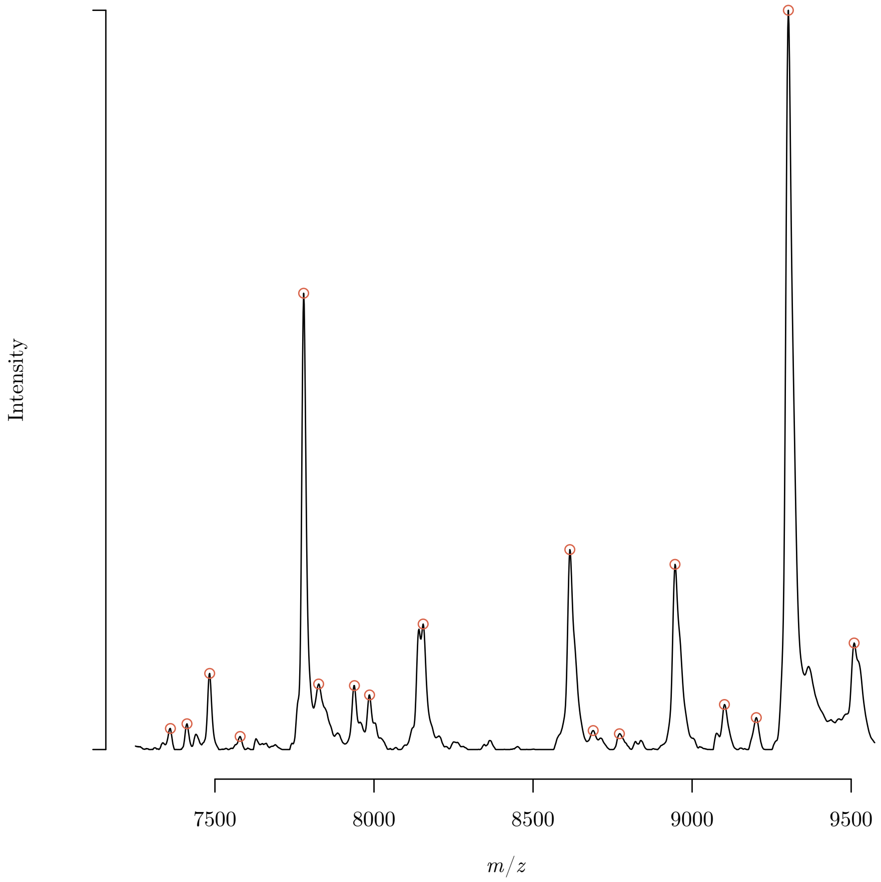


Figure 3.7: Simple S2N ratio peak detection on the Adam et al. (2002) data (patient 20) over a subset of the m/z-axis.

finding' algorithm can be applied to find local maxima of the negative signal. These are the troughs in the original signal. From this, peak width can be determined by the closest trough to the left and right of the identified peaks. This is a naive method to estimate peak widths as it assumes all peaks are fully resolved and without overlap.

If there is likely to be significant overlap of peaks, which hinders estimation of peak widths using the negative maxima method, a full-width at half-maximum (FWHM) type method can be employed. Such a method finds the width of the peak at half the height of the identified local maxima. Algorithmically, this is achieved by starting at the m/z-value of the local maxima and looking left and right to see when the signal falls below half the local maxima value. Much of the literature assumes Gaussian peaks (House et al., 2006; Yang et al., 2009; Barbarini and Magni, 2010) with height a_f and the peak shape taking the form $f(x) = a_f e^{-x^2/2\sigma^2}$. The relationship between the FWHM and the Gaussian standard deviation, σ, can be analytically deduced

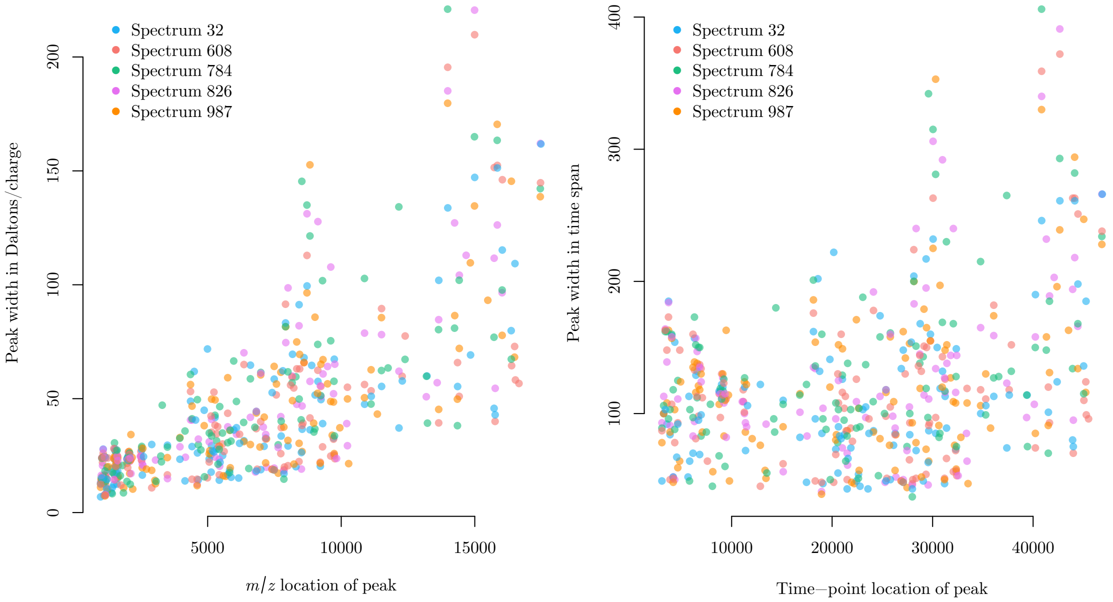


Figure 3.8: An illustration of peak width against peak location using TOF-values (left) and m/z-values (right). Five spectra from the GC mice dataset have been randomly selected to avoid unnecessary visual clutter. This information is only available after peak detection but illustrates that peaks are generally wider for larger molecular masses.

from

$$\frac{1}{2}f(0) = f(h_w),$$

where h_w is the width to the left or right of the peak's maximum a_f at $x = 0$. This results in FWHM $= 2\sqrt{2\ln 2}\sigma$, i.e.,

$$\begin{aligned} \frac{1}{2}a_f &= a_f e^{-h_w^2/(2\sigma^2)} \\ \Rightarrow \ln 2 &= h_w^2/2\sigma^2 \\ \Rightarrow h_w &= \pm\sqrt{2\ln 2}\sigma. \end{aligned}$$

Another method for determining peak width also uses Gaussian assumptions and aims to find an estimate of σ using non-linear regression:

$$\arg\min_{a,\sigma} \sum_{i=1}^{n_{h_w}} \left(y_i - f_{a,\sigma}(x_i)\right)^2,$$

where $f_{a,\sigma}(x_i) = ae^{-x_i^2/2\sigma^2}$ and x_i are m/z-values satisfying $f(x_i) \geq \frac{1}{2}a_f$. This method uses the x_i of the signal above the half maximum height. Alternatively, x_i satisfying $f(x_i) \geq p_h a_f$, where $p_h \in \left(0, \frac{1}{2}\right)$ could be used for an improved estimate of σ, as more of the peak shape is used in the estimation. Figure 3.9 shows the intensities that are greater than half the maximum intensity of peak 63 for patient 166 in the asthma2 dataset and the fitted non-linear regression line. The residuals from this fitted model do not appear to be random scatter. If fitting such non-linear regression to all peaks, many models will fail to converge because of non-Gaussian peak shape or insufficient separation from neighbouring peaks, even with sensible initial parameter estimates.

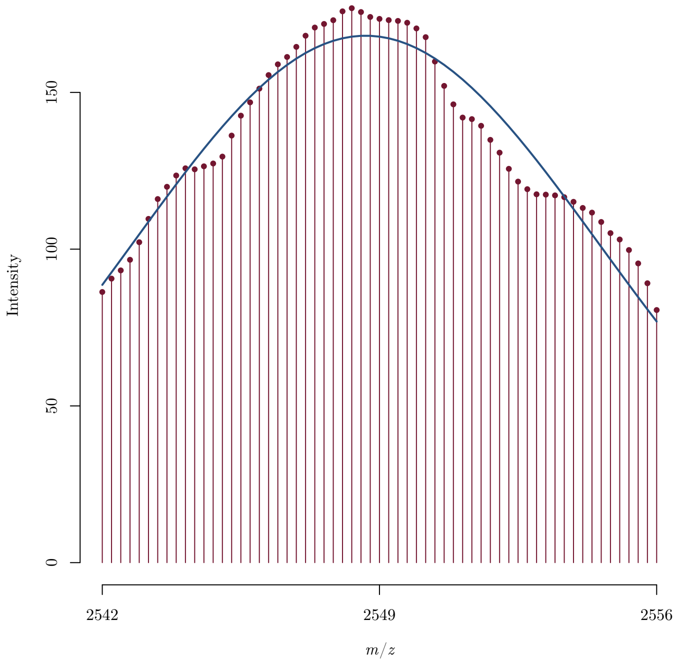


Figure 3.9: An estimated non-linear regression Gaussian curve (blue) fitted to the intensities (maroon) satisfying $f(x_i) \geq \frac{1}{2}a_f$.

The quantification of peak expression here in will be assumed to be peak height (maximum intensity) unless stated otherwise. Peak height is a standard treatment and is generally considered the most robust estimation of peptide quantity (Zhang et al., 2010). However, the different methods of quantifying expression outlined in this section will be compared with regard to which provides the strongest classification signal in Chapter 6.

3.3 Pre-processing step IV(b): peak alignment

Once peaks have been identified, it is important to identify peaks of the same biological origin across the spectra. Peak drift, or more generally, signal drift, is another noise component inherent in the spectra. Calibration of the TOF-system and conversion to m/z-values from retention time can cause non-linear drift of signal away from the true m/z-values. By performing peak alignment, comparison of peptide expression over all spectra is possible.

There are many algorithms that have been proposed for the purpose of peak alignment but the alignment method presented here allows for non-linear drift, is highly extensible and incorporates peak expression information in addition to peak m/z location by use of $(m/z, \text{intensity})$-pairs. The peak alignment method proposed here is based on algorithms outlined in Robinson et al. (2007). These algorithms are referred to as dynamic programming (DP), as they find subsequent optima of sub-problems. The algorithms in Robinson et al. (2007) were developed in the context of metabolomic gas chromatography-mass spectrometry (GC-MS) data and have required considerable modification for application to proteomic MALDI/SELDI TOF-MS data in the present thesis.

GC-MS produces a single spectrum for each chromatographic peak (which is simply an individual biological molecule) that is then hard-ionised, as opposed to the soft ionisation of MALDI/SELDI ('hard' meaning the molecule breaks into combinations of its constituent parts indicative of its chemical identity). Thus each spectrum is of the original chromatographic peak. In this way the GC-MS data are a matrix of spectra for each MS run as opposed to a vector of a single spectrum for each sample in MALDI/SELDI TOF-MS.

The following outlines the differences between GC-MS and MALDI/SELDI TOF-MS which impact on the alignment algorithms required.

- The data are different in structure; only a single vector of intensities is available for MALDI/SELDI TOF-MS per sample, as opposed to a matrix of spectra for GC-MS per sample. There is therefore more information to match (chromatographic) peaks in GC-MS than MALDI/SELDI TOF-MS peaks.
- The range of m/z values is far larger for MALDI/SELDI (with far larger molecular associated masses).
- Peak drift for MALDI/SELDI TOF-MS is related to the m/z values, while GC-MS data peak drift is considered on the (retention) time values.

Before explaining the modifications required for a dynamic programming approach of alignment to proteomic MALDI/SELDI TOF-MS, consider the problem of aligning two peak lists, L_1 and L_2, where L_1 is the list of peaks in spectrum 1 and L_2 is the list of peaks in spectrum 2. The peak lists contain information of peak quantity and location for each peak for either GC-MS or MALDI/SELDI TOF-MS. Let

$$L_1 = \begin{bmatrix} \boldsymbol{p}_1 \\ \boldsymbol{p}_2 \\ \vdots \\ \boldsymbol{p}_{n_1} \end{bmatrix} \text{ and } L_2 = \begin{bmatrix} \boldsymbol{q}_1 \\ \boldsymbol{q}_2 \\ \vdots \\ \boldsymbol{q}_{n_2} \end{bmatrix}.$$

Here $\boldsymbol{p}_i = (t_{p_i}, \boldsymbol{s}_{p_i})$ for the $i = 1, \ldots, n_1$ peaks in spectrum 1 and $\boldsymbol{q}_j = \left(t_{q_j}, \boldsymbol{s}_{q_j}\right)$ for the $j = 1, \ldots, n_2$ peaks in spectrum 2. Additionally $(t, \boldsymbol{s})$-pairs are the m/z and intensity information, respectively, of the peak found in the peak detection. Note the use of t to denote m/z-values; Robinson et al. (2007) naturally use t for retention time and this notation will be adopted here so as not to confuse the notation later on.

The aim is to match the $\boldsymbol{p}_i$, $\boldsymbol{q}_j$ pairs where $\boldsymbol{p}_i$ (or $\boldsymbol{q}_j$) do not necessarily match in a 1-1 fashion. For example, the true peak matching of lists L_1 and L_2 (that will be denoted $L_{1:2}$) may be

$$L_{1:2} = \left[\begin{array}{c:c} \boldsymbol{p}_1 & \bullet \\ \boldsymbol{p}_2 & \boldsymbol{q}_1 \\ \bullet & \boldsymbol{q}_2 \\ \boldsymbol{p}_3 & \boldsymbol{q}_3 \\ \vdots & \vdots \\ \boldsymbol{p}_{n_1} & \boldsymbol{q}_{n_2-2} \\ \bullet & \boldsymbol{q}_{n_2-1} \\ \bullet & \boldsymbol{q}_{n_2} \end{array}\right] \tag{3.8}$$

where $\bullet$ denotes ‘no-peak’.

This presents an optimisation problem with constraints that each peak must be matched with another peak (or no-peak) once and only once, and the peaks must be paired in chronological order (i.e. $\boldsymbol{p}_2$ cannot be matched before $\boldsymbol{p}_1$ is matched to another peak or no-peak).

A scoring system has been devised to create a metric to define what is in fact an ‘optimal’ peak-list match. One way of producing this optimisation is to use DP, and more specifically, maximum path algorithms. Maximum path algorithms work by constructing a peak similarity matrix of all combinations of peak pairings of the two peak lists, then finding a maximum scoring path through this similarity matrix where the path taken determines whether peaks are matched or not.

3.3.1 Dynamic programming

This section will briefly outline the Robinson et al. (2007) and Robinson (2008) DP GC-MS peak alignment algorithm to allow discussion of the methods and modifications required for successful peak alignment of MALDI/SELDI TOF-MS data.

Consider the GC-MS data and the peak similarity function, $P(\boldsymbol{p}_i, \boldsymbol{q}_j)$, which provides a similarity scoring for any peak pairing $\boldsymbol{p}_i$, $\boldsymbol{q}_j$ that takes into account both the peak location similarity and the peak profile (intensities).

The elements in the peak similarity matrix, $P = [P_{i,j}] = \left[P(\boldsymbol{p}_i, \boldsymbol{q}_j)\right]$, take the form

$$P(\boldsymbol{p}_i, \boldsymbol{q}_j) = S(\boldsymbol{s}_{p_i}, \boldsymbol{s}_{q_j})T(t_{p_i}, t_{q_j}), \tag{3.9}$$

where S is the signal similarity and T is the time or location similarity.

The S used is the normalised dot product of the intensity values (cosine angle between the vectors),

$$S(\boldsymbol{s}_{p_i}, \boldsymbol{s}_{q_j}) = \frac{\boldsymbol{s}_{p_i} \cdot \boldsymbol{s}_{q_j}}{||\boldsymbol{s}_{p_i}||\;||\boldsymbol{s}_{q_j}||} = cos\theta, \qquad \text{and} \tag{3.10}$$

$$T(t_{p_i}, t_{q_j}) = e^{\frac{-(t_{p_i} - t_{q_j})^2}{2D^2}}, \tag{3.11}$$

where D is a constant predefined to penalise retention times that are increasingly distant. Also note, $\boldsymbol{u} \cdot \boldsymbol{v} = u_1v_1 + u_2v_2 + \ldots + u_nv_n$ and $||\boldsymbol{u}|| = \sqrt{u_1^2 + u_2^2 + \ldots + u_n^2}$. It is also worth noting that $S(\boldsymbol{s}_{p_i}, \boldsymbol{s}_{q_j}) \in [0, 1]$ and $T(t_{p_i}, t_{q_j}) \in (0, 1]$, therefore $P(\boldsymbol{p}_i, \boldsymbol{q}_j) \in [0, 1]$.

With a matrix of peak similarities constructed, P, a maximum path algorithm needs to be used to find an optimal alignment of the peaks. The maximum path algorithm used in Robinson et al. (2007) is the Needleman and Wunsch (1970) algorithm of global alignment. Originally proposed as a nucleotide sequence matching algorithm (i.e. matching of discrete values or letters), it is used successfully on continuous variables taking values from 0 to 1, as is the case here.

The Needleman and Wunsch (NW) algorithm creates a new matrix,

$$M = [M(i, j)]_{(n_1+1)\times(n_2+1)},$$

defined recursively for $i = 0, 1, \ldots, n_1$ and $j = 0, 1, \ldots, n_2$,

$$M(i, j) = \max \left\{ \begin{array}{l} M(i-1, j-1) + P(p_i, q_j) \\ M(i-1, j) - \delta_G \\ M(i, j-1) - \delta_G \end{array} \right\}, \tag{3.12}$$

with boundary values $M(i, 0) = -i\delta_G$ and $M(0, j) = -j\delta_G$. The gap penalty, $\delta_G \geq 0$, needs to be chosen carefully or calibrated.

M is created by initialising the top row and left column with negative values increasing in size by position. Then element $(1, 1)$ is evaluated using Equation (3.12). $M(1, 1)$ will of course be assigned the value $M(i-1, j-1) + P(p_i, q_j) = M(0, 0) + P(p_1, q_1) = 0 + P(p_1, q_1) = P(p_1, q_1)$ for non-negative values for the function P. This recursive assessment of the values of M can be continued along row 1 and column 1 before assigning the value $M(2, 2)$ and following a similar pattern. Once all the values of M are evaluated, the algorithm works backwards (in reference to evaluating the elements in M) from $M(n_1, n_2)$. The decision to move up, left or diagonally up to the left determines the alignment. Moving up corresponds to matching peak $\boldsymbol{p}_{n_1}$ to 'no-peak', moving left corresponds to matching peak $\boldsymbol{q}_{n_2}$ to 'no-peak' and a move diagonally corresponds to matching peaks $\boldsymbol{p}_{n_1}$ and $\boldsymbol{q}_{n_2}$. The decision to make one of these three moves is dependent entirely on which of $M(n_1 - 1, n_2), M(n_1, n_2 - 1)$ and $M(n_1 - 1, n_2 - 1)$ are largest, respectively. Once the algorithm moves to a new element the next move is similarly recursively made. By traversing the M matrix until row 0 or column 0 is reached, a path has been created which maximises the score of the alignment. This path is interpreted to assign matched peaks according to the up, left or diagonal movements.

In the traversal of M, if two or three adjacent cells that are up, left or diagonal contain exactly the same values, this suggests there are multiple alignments which achieve maximum scoring for the algorithm. This event is very unlikely in the scenario of GC-MS (or TOF-MS generally) as the values in P are continuous. If the path reaches row 0 or column 0 in a position that is not $M(0, 0)$, this simply means that sequence 2 or sequence 1 have peaks matched to no-peaks, respectively.

The alignment of two peak lists has been explained, but of course there are likely to be tens to thousands of samples to be aligned. It is proposed in Robinson et al. (2007) that alignments be successively amalgamated (in a pairwise fashion). Robinson et al. (2007) refers to N-M alignments where there are N samples previously aligned in the first peak alignment and M samples in the second alignment. Note the peaks lists previously mentioned are simply 1-alignments and the resulting alignment in Equation (3.8) is a 2-alignment. The N-alignment and M-alignment, respectively, take the form

$$L_N = \begin{bmatrix} \boldsymbol{p}_{11} & \boldsymbol{p}_{12} & \cdots & \boldsymbol{p}_{1N} \\ \boldsymbol{p}_{21} & \boldsymbol{p}_{22} & \cdots & \boldsymbol{p}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{p}_{K1} & \boldsymbol{p}_{K2} & \cdots & \boldsymbol{p}_{KN} \end{bmatrix} \quad \text{and} \quad L_M = \begin{bmatrix} \boldsymbol{q}_{11} & \boldsymbol{q}_{12} & \cdots & \boldsymbol{q}_{1M} \\ \boldsymbol{q}_{21} & \boldsymbol{q}_{22} & \cdots & \boldsymbol{q}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{q}_{L1} & \boldsymbol{q}_{L2} & \cdots & \boldsymbol{q}_{LM} \end{bmatrix},$$

where $\boldsymbol{p}_{ia}$ is the peak information of the i^{th} peak for the a^{th} sample in the N-alignment. Note, $\boldsymbol{p}_{ia}$ may be empty for peaks not found for the a^{th} sample at the

i^{th} peak in the N-alignment but at least one $\boldsymbol{p}_{ia} \neq \bullet$ for $a = 1, \ldots, N$ for each i. The $\boldsymbol{q}_{jb}$ for $b = 1, \ldots, M$ of the M-alignment are similarly defined. The similarity matrix for peaks of two alignments, that may have been constructed from previous alignments is calculated as the average of peak similarities of the peaks (row) i in the N-alignment and peaks (row) j in the M-alignment from the samples within the alignments using

$$W(i,j) = \frac{\sum_{a=1}^{N}\sum_{b=1}^{M} P\left(\boldsymbol{p}_{ia}, \boldsymbol{q}_{jb}\right)}{\sum_{a=1}^{N}\sum_{b=1}^{M} I\left[P\left(\boldsymbol{p}_{ia}, \boldsymbol{q}_{jb}\right) > 0\right]},$$

with $P(\cdot)$ defined in Equation (3.9) with the additional case of $P\left(\boldsymbol{p}_{ia}, \boldsymbol{q}_{jb}\right) = 0$ when $\boldsymbol{p}_{ia}$ or $\boldsymbol{q}_{jb}$ have no peak.

Using W, the same process of finding the maximum path alignment is undertaken. Alignment using W can simply be seen as a generalisation of the alignment created by P and $W = P$ for a 2-alignment. In this way, successive alignments can be amalgamated to align all the samples. For example, the alignment of peak lists L_N and L_M may look like

$$L_{N:M} = \left[\begin{array}{cccc:cccc} \boldsymbol{p}_{11} & \boldsymbol{p}_{12} & \cdots & \boldsymbol{p}_{1N} & \boldsymbol{q}_{11} & \boldsymbol{q}_{12} & \cdots & \boldsymbol{q}_{1M} \\ \boldsymbol{p}_{21} & \boldsymbol{p}_{22} & \cdots & \boldsymbol{p}_{2N} & \bullet & \bullet & \cdots & \bullet \\ \bullet & \bullet & \cdots & \bullet & \boldsymbol{q}_{21} & \boldsymbol{q}_{22} & \cdots & \boldsymbol{q}_{2M} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \boldsymbol{p}_{K1} & \boldsymbol{p}_{K2} & \cdots & \boldsymbol{p}_{KN} & \boldsymbol{q}_{L1} & \boldsymbol{q}_{L2} & \cdots & \boldsymbol{q}_{LM} \end{array}\right].$$

Figure 3.10 shows the W matrix of two 4-alignments for the de Noo et al. (2006) dataset that would in turn create an 8-alignment.

The final question of such a DP method of alignment is how to choose the order of the successive alignments. Robinson et al. (2007) creates the statistic $T_{N:M} = \sum_k z_k$, where k is the index of the alignment between lists L_N and L_M (i.e. the peak number or row number in the resulting alignment) and z_k is defined as,

$$z_k = \begin{cases} W(i,j) & \text{if } k \text{ is a position where } i,j \text{ are matched} \\ -\delta_G & \text{if } k \text{ is a position where } i,j \text{ are not matched.} \end{cases}$$

The metric, $T_{N:M}$, indicates the similarity of two peak lists. Robinson et al. (2007) then finds the $T_{N:M}$ for every pairwise alignment of all samples to create a distance matrix between samples to guide the order of amalgamation of samples. The creation of a dendrogram (guide tree) from this distance matrix allows the successive amalgamation of peak lists until a peak alignment of all samples is created.

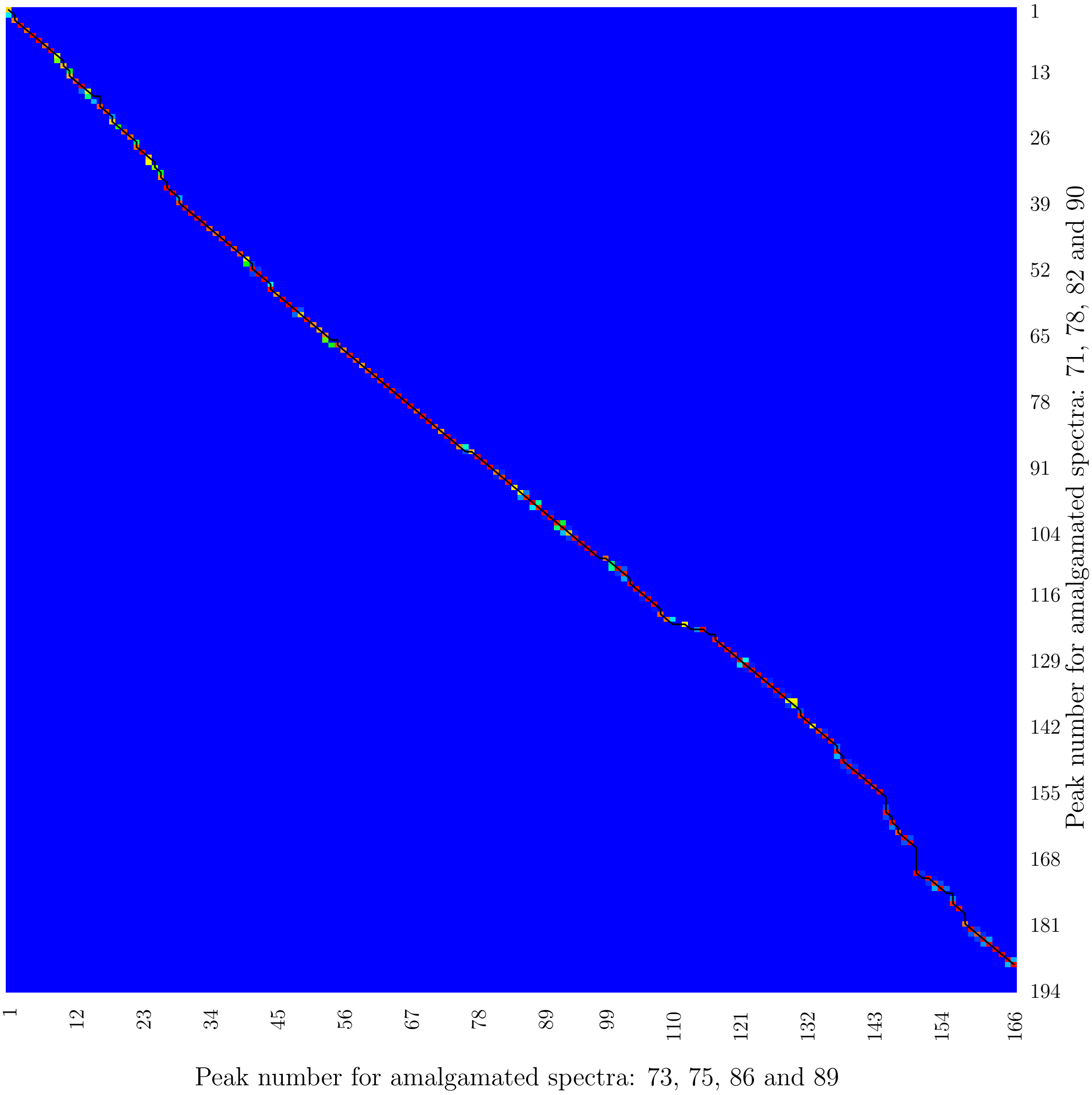


Figure 3.10: W matrix for the amalgamated alignment of two 4-alignments of the de Noo et al. (2006) dataset.

3.3.2 Modifications required for MALDI/SELDI TOF-MS

The following modifications were required to implement the Robinson et al. (2007) alignment method for MALDI/SELDI TOF-MS data. As non-trivial changes were required, an implementation of the new alignment methods as `R`-functions is provided in Appendix A.3.

(1) The T function needs to be changed to allow for increasing mass drifts for increasing m/z-values.

(2) The multiplicative relationship between the S and T functions to produce P needs to be altered, as intensity information is not as important as (m/z) location information in the MALDI/SELDI TOF-MS setting.

(3) As chromatographic peaks are not produced in MALDI/SELDI TOF-MS, different S functions need to be considered.

(4) A different maximum path algorithm needs to be used to account for larger dissimilarity between spectra than exists for GC-MS.

(5) The calculation of the guide tree for amalgamation of alignments needs to be changed for computational efficiency because of the large number of spectra and peaks in the available MALDI/SELDI TOF-MS data.

Modification (1) was addressed by redefining T from Equation (3.11) as

$$T(m_{p_i}, m_{q_j}) = e^{\frac{-(m_{p_i}-m_{q_j})^2}{2\left(D\min\left\{m_{p_i}, m_{q_j}\right\}\right)^2}}, \tag{3.13}$$

where m_{p_i}, m_{q_j} are the m/z-values of peaks $\boldsymbol{p}_i, \boldsymbol{q}_j$ respectively.

Equation (3.13) now accounts for peak drift proportional to m/z as the TOF-MS literature predicts (Coombes et al., 2005; Orvisky et al., 2006). Equally, $(m_{p_i}+m_{q_j})/2$ could be used in the exponent in the denominator as opposed to $\min\left\{m_{p_i}, m_{q_j}\right\}$, but the latter is more conservative. The interpretation of D is the allowable proportion of drift considered plausible for any mass value as opposed to the maximum allowable retention time difference.

In Robinson et al. (2007) the peak similarity, P, is generated by multiplying the intensity similarity, S, and the location similarity, T, effectively giving both functions S and T equal weight. This is acceptable in a LC-MS setting where the information provided by both functions could be seen as equally important. In the MALDI/SELDI TOF-MS setting, however, equal importance is not likely to be the case. Proposed here for Modification (2) is a peak similarity function to take the form,

$$P\left(\boldsymbol{p}_i, \boldsymbol{q}_j\right) = \lambda S\left(\boldsymbol{s}_{p_i}, \boldsymbol{s}_{q_j}\right) + (1-\lambda)\, T\left(m_{p_i}, m_{q_j}\right), \tag{3.14}$$

where $\lambda \in [0, 1]$ is a constant to weight the S and T functions. A discussion about the choice of λ will be addressed with Modification (3). The function P could also be defined as $P\left(\boldsymbol{p}_i, \boldsymbol{q}_j\right) = S\left(\boldsymbol{s}_{p_i}, \boldsymbol{s}_{q_j}\right)^{\lambda} T\left(m_{p_i}, m_{q_j}\right)^{\frac{1}{\lambda}}$ but Equation (3.14) is more appealing for the current setting. Even if the peak shapes are very different but the mass locations are exactly the same, the flexibility of an additive relationship is required.

Modification (3) is required because MALDI/SELDI TOF-MS data do not have associated chromatograms for each peak to calculate the proposed S function dot-product in Robinson et al. (2007). Peak shape and intensity metrics to compare peak similarity were used instead.

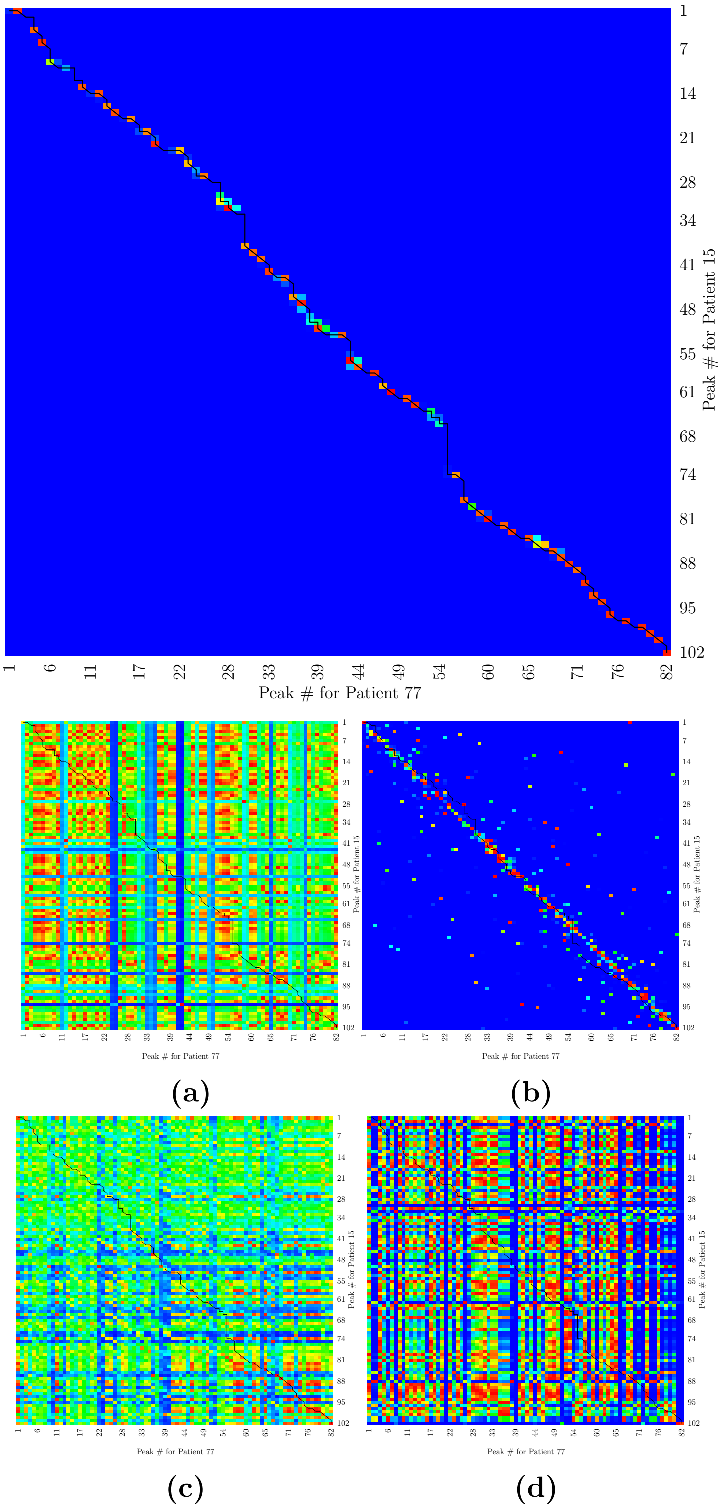
The simplest function for comparing peak intensity (but not shape) is to create relative differences in maximum peak intensity. It was hoped this would yield a similarity matrix that reflected a scenario where peaks at a similar TOF would have a similar intensity, i.e. peaks that are of the same protein would be roughly similar in intensity across spectra. However, this was not observed with enough consistency to effectively aid the peak alignment process; the noise far outweighed the similarities. Figure 3.11 illustrates experimental examples of similarity matrices with different signal similarity functions.

The second method considered for generating a signal similarity matrix, S, was similar to the GC-MS method of Robinson et al. (2007) and Robinson (2008). A normalised dot product of peak intensities was used. This required a pre-specified number of intensity points to the left and right of detected peaks for all peaks to be defined. As has been discussed throughout this thesis, peaks are generally wider at larger time points, so such a method is problematic. From Figure 3.11 it can be seen the S matrix created using the normalised dot product of the intensities around the detected peaks is overwhelmingly noise.

Another method considered to extract peak shape and signal similarity fitted a Gaussian curve to each detected peak (similar to that of §3.2),

$$f(x) = \hat{a} e^{\frac{-(x-\hat{b})^2}{2\hat{\sigma}^2}}, \tag{3.15}$$

where $\hat{a}, \hat{b}, \hat{\sigma}$ are non-linear regression parameter estimates of the peak parameters a, b, σ respectively. The underlying assumption here is that a protein's expression across samples may be different but the shape of the peak (as a result of the inability of the mass spectrometer to resolve the individual isotopic peaks) of the protein should be the same. This is why broadened peaks for larger mass-values are resolved: there are more possible isotopic versions of the protein. The parameter a is an amplitude parameter and is of no use in comparing peak shapes and b is a m/z offset parameter which is not useful in comparing peak shape either. However, σ is the standard deviation of the Gaussian curve and is a representation of the peak's



(a)

(b)

(c)

(d)

Figure 3.11: A calibrated T matrix (top) for two spectra from the Adam et al. (2002) data. The black line indicates the true peak alignment path via observation of the spectra. The remaining four heatmaps below are S matrices using the metrics: (a) relative differences in maximum peak intensity, (b) modified version of (a) weighted for peak location, (c) normalised vector dot product, and, (d) peak width.

width or isotopic distribution. This was not a successful method however as the resulting S matrix was observed to be noise (Figure 3.11(d)).

Since an informative signal similarity matrix could not be found, the choice to employ the weighting $\lambda = 0$ for the signal similarity from Equation (3.14) was logical. Although this means the peak similarity matrix becomes the location similarity matrix, if a suitable signal similarity function could be found, this would provide another advantage of the DP method of alignment as it could incorporate the additional information of peak shape.

In initial testing of the NW algorithm via the use of the `R` package `flagme` (Robinson, 2010) showed unsatisfactory alignments which required Modification (4). An assumption with global alignment algorithms is that the sequences are sufficiently similar. A local alignment algorithm, Smith and Waterman (1981), is a variant of the original algorithm that modifies the objective of global alignment to allow sequences that have more distant similarity to be aligned successfully. The Smith and Waterman (SW) algorithm does not assume the sequences should be roughly matched at the beginning and end. A modified version of the SW algorithm proposed here is also recursive and is as follows:

$$M(i,j) = \max \left\{ \begin{array}{l} I\left[P(p_i, q_j) < \tau_G\right] \left(M(i-1, j-1) + P(p_i, q_j)\right) \\ M(i-1, j) - \delta_G \\ M(i, j-1) - \delta_G \end{array} \right\} \tag{3.16}$$

where $\delta_G \geq 0$ is the gap penalty that needs to be carefully calibrated as for the NW algorithm, I is the indicator function $\in \{0, 1\}$, $\tau_G \in [0, 1]$ is a threshold peak similarity and $M(i, 0) = M(0, j) = 0$. The different initialisation of $M(i, 0) = M(0, j) = 0$ compared to the NW algorithm is because the algorithm is not trying to encourage alignment from the top-left corner of M to the bottom-right corner of M as in global alignment, but rather, trying to find maximum scoring sub-alignment anywhere in M. The inclusion of τ_G discourages the match of $\boldsymbol{p}_i, \boldsymbol{q}_j$ if their peak similarity is not sufficiently large.

The same interpretation of the NW algorithm can be applied to the SW algorithm in terms of finding the maximum path; the backtracking of the maximum path matrix in the directions up, left or diagonally up to the left have the same interpretation but M is computed slightly differently. For the local alignment, the path finding does not start at element $M(n_1, n_2)$ as before but rather the maximum value in M. From there, the matrix is traversed as previously until an element of 0 is chosen. This means the algorithm halts and the maximum scoring sub-alignment has been found. A naive implementation of the SW algorithm[1] was converted to `R`-code with the modifications of (3.16).

[1] `C`-code can be found at: https://code.google.com/p/swalign/

Modification (5) was considered as calculation of $T_{N:M} = \sum_k z_k$ for every possible pairwise 2-alignment can be time-consuming for upwards of several hundred or even thousands of spectra and T, S, P and M matrices need be evaluated as well as the maximum path before $T_{N:M}$ can be calculated.

The guide tree to determine the amalgamation order of alignments can be alternatively created based on a matrix of pairwise distances between expression profiles of spectra. A predefined distance metric can be computed on the pre-processed spectra expression values and a dendrogram, or guide tree, can be generated from the distance matrix using a predefined linkage metric. Spectra with more similar expression profiles as computed by the distance metric will be aligned together earlier in the guide tree amalgamation order. From experimentation, the Euclidean distance metric and average linkage for the dendrogram provided the most sensible amalgamation order similar to that dictated by $T_{N:M}$; an example amalgamation dendrogram can be seen in Figure 3.12.

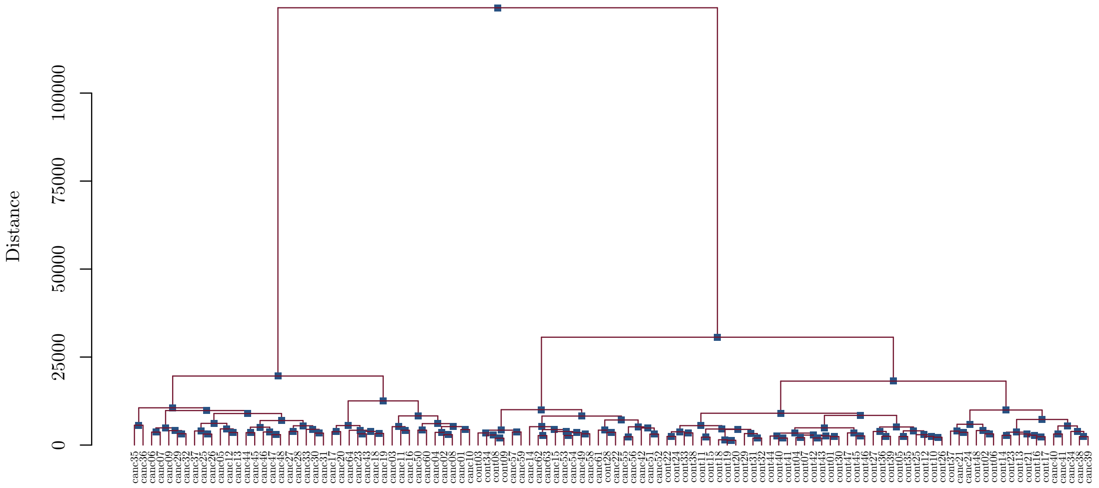
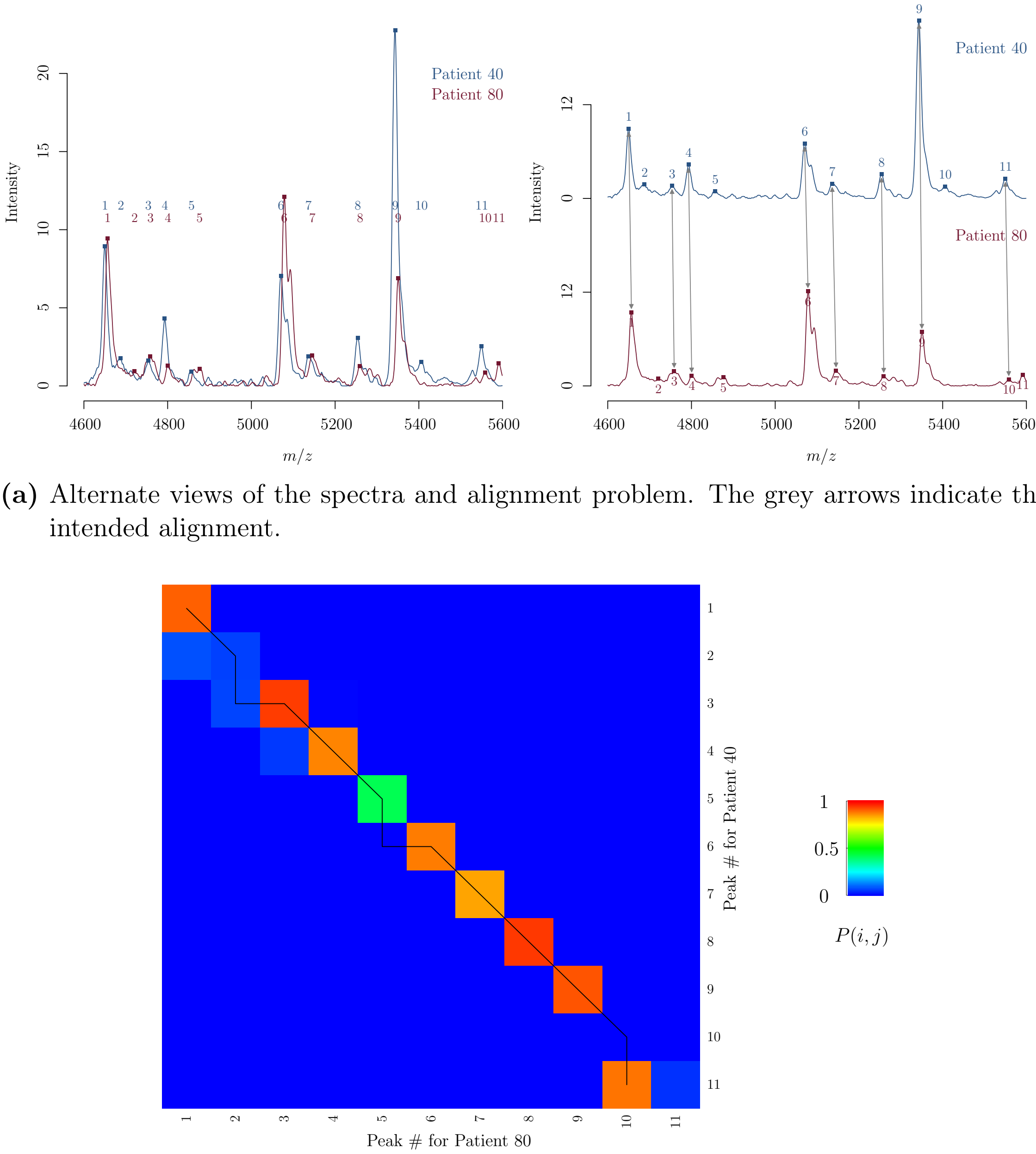


Figure 3.12: Alignment amalgamation de Noo et al. (2006) spectra.

Pairwise example for proteomic MALDI/SELDI TOF-MS data

Figure 3.13 provides a scaled example of the process required to align two peak lists in a pairwise fashion using the DP alignment algorithm outlined, modified for MALDI/SELDI TOF-MS data. Here a subset of the m/z-axis of the spectra for patients 40 and 80 was used from the Adam et al. (2002) dataset.



(a) Alternate views of the spectra and alignment problem. The grey arrows indicate the intended alignment.

(b) A peak similarity matrix and maximum path (black line) enabling correct alignment of the peaks.

Figure 3.13: Peak alignment example for two spectra along a subset of the m/z-axis.

3.3.3 Implementation and calibration

The DP method for MALDI/SELDI TOF-MS peaks outlined in §3.3.2 dictates that a set of spectra can be aligned by:

(1) Calculating pairwise spectra similarity statistics.

(2) Creating a dendrogram of spectra alignments based on similarity measures.

(3) Performing dynamic programming of N- and M-alignments of spectra based on the dendrogram amalgamation until all alignments have become one N-alignment.

However, the DP alignment approach requires optimised choices of the parameters D, δ_G and τ_G. A summary of these parameters for reference is given in Table 3.3. D and τ_G are related, as an increase in D will increase mass similarities and the related threshold, τ_G, in which a match of peaks is acceptable. δ_G is the gap penalty of the alignment: the more similar the spectra peaks are expected to be, the larger this parameter should be. Note that the calibration of λ is overlooked here as no suitable relationship between peak intensities was found as per Modification (3) in §3.3.2. Consequently, λ was set to 0.

Table 3.3: Dynamic programming parameters.

Parameter	Values/Range	Explanation
D	$(0, 0.003]$	Roughly, a proportion of relative mass-drift acceptable between peaks.
δ_G	$(0, 1)$	The gap penalty in the max path algorithm.
τ_G	$(0, 1)$	The minimum allowable peak similarity to warrant a match.

Two approaches were trialled to optimise DP alignment parameters. The first approach required the creation of pairwise 'truth' peak alignments so the accuracy of potential DP alignments could be tested. The truth alignments were created by visual inspection of pairwise spectra. Then estimated DP alignments for given input parameters were then compared to the truth alignment via a metric of 'correctness'. The metric to quantify alignment correctness was chosen here to be the harmonic mean (HM) that has been used in similar situations (Kim et al., 2011). The HM provides a robust metric where the number of true peak matches and the number of estimated peak matches are not assumed to be equal, and importantly, both the number of correctly identified peak matches and the number of incorrectly identified peak matches influence the metric's value. The HM is given by

$$\mathrm{HM} = \frac{\mathrm{TP}}{\left(n_{\mathrm{truth}} + \hat{n}_{\mathrm{truth}}\right)/2},$$

where TP is the number of true positives predicted by the peak matching algorithm, n_{truth} is the number of true peak pairings, and $\hat{n}_{\text{truth}}$ is the estimated number of peak pairings. Note that TP $\leq \min\left(n_{\text{truth}}, \hat{n}_{\text{truth}}\right)$ and therefore HM $\in [0, 1]$. HM $= 1$ is obtained only when perfect alignment has occurred. Peak matches refer to peaks being matched to another peak or no-peak.

The parameter set $\{D, \delta_G, \tau_G\}$ can be optimised to produce a maximised HM over a sample of pairwise alignments and then the alignment of all spectra can be run using the optimised parameters. A grid search approach can be undertaken and the log-odds of HM can be modelled by an over-specified (to account for the unknown functional form and relationship between the outcome and predictors) linear mixed effects model (LME, to factor in the dependence structure between observations as spectra alignments are repeated). The LME can then be reduced in complexity until a 'best' model is reached. The best model objective can be decided by the Bayesian Information Criterion. As the resulting model will be a multivariate polynomial function (bounded), an estimated optima, $\widehat{\text{HM}}$, using an initial condition can be found.

A modelled and optimised dynamic peak alignment using the LME approach was initially undertaken but it became apparent that extending this method to amalgamations of alignments beyond the pairwise case would be prohibitively time consuming when creating truth peak matches for 3-alignments, 4-alignments, up to the final n-alignment. It is also unclear how the N-alignments should be sampled and respective information fully incorporated into the LME model.

A second, more practical approach to optimising the DP alignment parameters was then employed. To best visualise multiple spectra and alignments, heatmaps were used. The heatmaps represent spectra as rows, m/z-values as columns and intensity as heat colouring. Overlaid on the heatmap are detected peaks and the subsequent alignment peak number (with colouring for identification). From such a plot it is possible to gauge whether the alignment parameters are performing sensible alignments. In practice, visual parameter optimisation can be performed on a random selection of spectra for efficiency. Experience suggests optimised parameters for a randomly selected subset of spectra is generally indicative of the optimised parameters for the alignment of the entire set of spectra.

An example heatmap, while optimising alignment parameters on a randomly selected subset of GC mice spectra, can be seen in Figure 3.14. Rows represent spectra where group membership of each spectrum is denoted by the group colour on the left. The columns are m/z-values and the respective intensities for each spectrum are illustrated by greyscale; darker greys indicate greater intensities up to 6229 units. Additionally, black strips indicate a detected peak and the aligned peak number with common colour is overlaid.

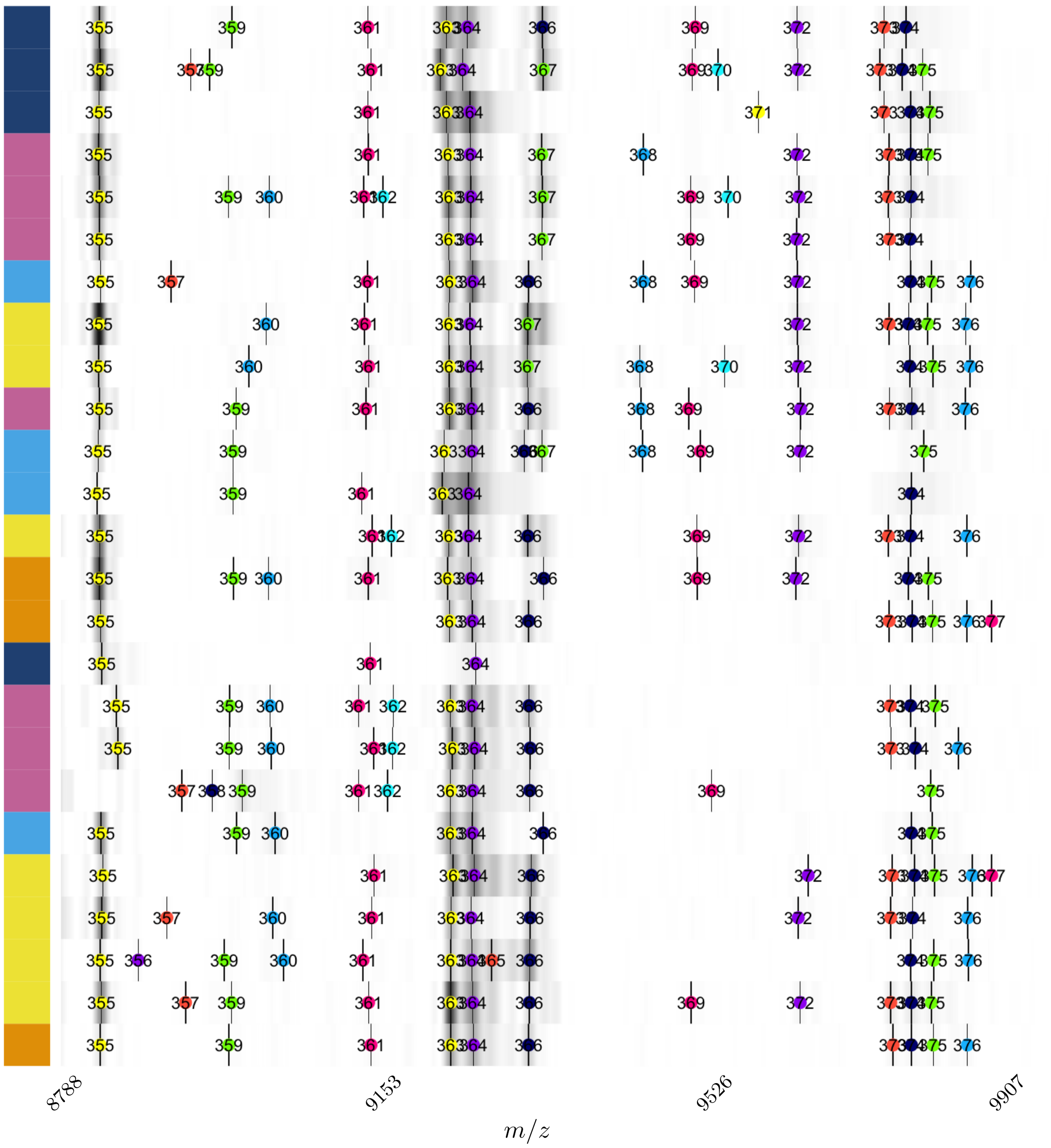


Figure 3.14: A subset of the GC mice spectra over the 8788-9907Da subinterval of the m/z domain; a potential alignment.

3.3.4 Comparison to standard peak alignment techniques

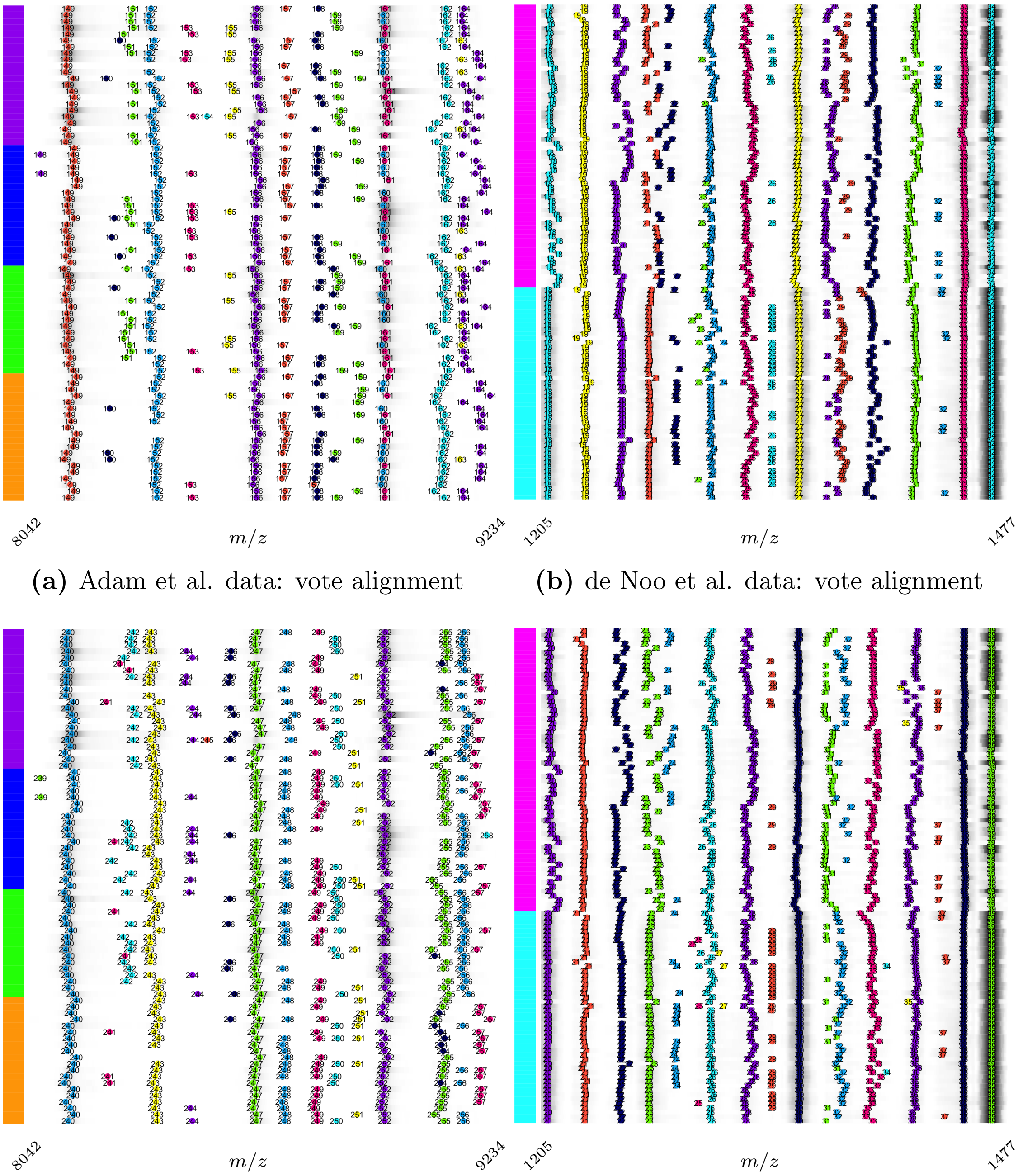
The DP alignment is compared to the iterative heuristic method proposed in Yasui et al. (2003) and Adam et al. (2002), widely used in the GC-MS and TOF-MS literature (Wong et al., 2005; Mitchell et al., 2005; Kazmi et al., 2006) and implemented in `msProcess` (Gong et al., 2012). For labelling purposes, this alignment method will be referred to as the vote method.

The vote method requires a parameter analogous to the parameter D of the DP alignment of §3.3.2, which is an acceptable drift of peaks as a proportion of their location. Using this parameter, each m/z-value is assigned the number of peaks that have been detected within an acceptable region of drift for that m/z-point. The m/z-value deemed to have the most peaks in the drift region is assigned as the location of the peaks in the drift window; the peaks are removed as 'aligned' peaks and the process is repeated until all peaks are aligned.

Although not pursued for greater than pairwise alignment (as discussed in §3.3.3), Table 3.4 illustrates the optimised alignment with respect to HM for both DP and vote alignment on the Adam et al. (2002) data. Random numbers were used to select pairs of spectra, for which truth peak matches were created by visual inspection. The two alignment methods were optimised by grid search, and the maximum HM is reported in Table 3.4. Unfortunately, generalising beyond pairwise alignment to multiple alignment from this table would be tenuous. The pairwise alignment properties cannot be guaranteed to hold for multiple alignments and each pair of spectra in Table 3.4 was optimised individually, whereas multiple alignments are made using fixed alignment parameters. Despite the limitations in interpretation of the HMs in Table 3.4, the DP alignment was slightly more favourable and is consistent with the generalised alignment seen in Figure 3.15.

Table 3.4: Harmonic means of pairwise dynamic programming and vote alignment on randomly sampled Adam et al. (2002) spectra.

Alignment		HM (%)		
Spectrum$_1$	Spectrum$_2$	DP	Vote	Difference
15	77	97.3	98.9	-1.5
40	80	95.6	91.3	4.3
87	151	98.6	95.0	3.6
97	143	100	97.2	2.8
171	199	100	98.6	1.4
169	187	96.9	95.9	1.0
256	267	94.2	94.7	-0.6
296	320	100	100	0
Mean		97.8	96.5	1.4



(a) Adam et al. data: vote alignment

(b) de Noo et al. data: vote alignment

(c) Adam et al. data: DP alignment

(d) de Noo et al. data: DP alignment

Figure 3.15: Alignment heatmaps of two datasets and two alignment methods for comparison.

Figure 3.15 depicts spectra alignment using the two alignment methods discussed on two datasets, Adam et al. (2002) and de Noo et al. (2006). Subintervals of the m/z-axis were chosen where a substantive proportion of the peaks lay for the spectra. It is clear neither method produces alignments that would be considered by visual inspection to be perfectly correct. Especially for the Adam et al. (2002) data, the DP alignment is superior. The difference in methods is apparent as the DP alignment can be thought of as a soft or fuzzy alignment, as the peak similarities can take a continuous range of values, while the vote method creates a harder alignment, only considering peaks in defined windows. This property of the vote method can cause strange separations between peaks of the same biological origin that should be aligned together into two or more groups of peaks. This can be attributed to the allowable drift windows of abundant peaks 'catching' different peaks nearby that have slightly more drift towards the most abundant peaks. An example of this phenomenon are the peaks at approximately 9000Da of the Adam et al. (2002) spectra in Figure 3.15.

Robinson et al. (2007) noted that agglomeration ordered alignment was a sensible approach to make the pairwise alignment problem computationally tractable. Research into agglomeration orders, or other methods, to create N-alignments may further improve the DP alignment sensitivity but is not pursued here because sensible alignments were achieved using the method discussed. The DP alignment has the advantage of flexibility over the vote method, especially in allowing non-symmetrical drifts of peaks in both directions from the true mass location. The DP approach also is extensible, especially if signal similarities can be harnessed in addition to mass similarity. A successful signal similarity metric was not found here despite experimentation. This could be an area of further research but is potentially limited by the current reproducibility issues of the MALDI/SELDI TOF-MS technology.

3.4 Recommendations

Normalisation: normalisation is a pre-processing step that may influence the downstream analyses. It is clear TCN is a naive approach to normalisation that cannot fully account for experimental bias. EQN, as proposed here, is a simple and effective normalisation technique that reduces noise for differentially expressed proteins. CLN is a normalisation method that is far more computationally intensive, or even prohibitive, that has inferior results to EQN. Normalisation, like all pre-processing steps, can be significantly aided by information and graphics depicting batch-order information to assess the effectiveness of the chosen normalisation method.

Peak alignment: an alternative method of peak alignment is proposed in this thesis, namely dynamic programming, drawing on the GC-MS literature. It is compared to the standard voting and clustering method of alignment and performs equally as well if not better with its more flexible structures and a pseudo-probabilistic basis. R code is provided so further research on its robustness and generally applicability can be assessed on future datasets.

Chapter 4

Data visualisation, intermediate analysis and biomarker discovery

This chapter presents the initial analysis of the peak expression data obtained from the spectra pre-processing. Visual analysis of the peak expression data allows investigation of experimental effects such as batch effects and biological signal with regards to disease group classification. The important process of identifying potential biomarkers for further research requires appropriate statistical modelling. Linear models are employed to identify peaks, and hence peptides, which have differing expression levels between disease groups, while controlling for experimental factors. The missingness that occurs in MALDI/SELDI TOF-MS peak expression data may not be ignorable. Linear models are fitted on both data with missing values and data with imputed values to investigate whether the assumption of values missing at random is valid. Finally, surrogate variable analysis and remove unwanted variation are evaluated for their utility in analysing MALDI/SELDI TOF-MS peak expression data.

4.1 Unsupervised exploratory analysis

The pre-processing of the MS data transforms the data from raw, incomparable spectra to a concise set of information, namely a matrix, Y, that consists of peak expressions. The primary aim of finding relationships between the peptides and disease states can now be explored.

Exploratory analysis of the data to investigate patterns or relationships between variables can be performed using *unsupervised* learning. In this context, unsupervised learning creates groupings of the data without knowledge of an observation's membership to a disease group. This is in contrast to *supervised* learning where the analysis utilises the disease group membership as an outcome variable. Supervised learning for diagnostic purposes is addressed in Chapters 5 and 6.

4.1.1 Principal component analysis

Principal component analysis (PCA) is a standard method of unsupervised learning conceived early in the twentieth century (Pearson, 1901). PCA uses the entirety of the peak expression data, without the knowledge of group labels or experimental factors, and finds linear combinations of the peak expressions that maximise the variability, estimated by the sample covariance matrix $S_Y = \frac{1}{n}Y^TY$, across the observations (Jolliffe, 2005). Note the observed data Y are assumed to be mean centred and scaled as PCA is not transform invariant. PCA also assumes 'complete' data, the occurrence of missing data is addressed in the next section.

Geometrically, PCA is a transformation of the observations $\boldsymbol{y}_i$ to corresponding transformed $\boldsymbol{z}_i$, via a new orthonormal basis that is optimised to maximise the variance of the observations. Such a transformation is constructed using eigenvectors and eigenvalues (Shlens, 2009). The transformed data are generated through the relation $Z^T = \Omega^TY^T$, where Y is an $n \times P$ matrix and $\Omega = [\boldsymbol{\omega}_1\, \boldsymbol{\omega}_2 \ldots \boldsymbol{\omega}_P]$ is a $P \times P$ matrix with the columns corresponding to the P (right) eigenvectors of the sample variance, S_Y. The P eigenvectors and eigenvalues satisfy $S_Y\Omega = \Omega\Lambda$, where Λ is a diagonal matrix of the ordered eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_P$, descending in value (Jolliffe, 2005).

The eigenvalue λ_j represents the proportionate amount of variability that the principal component j provides. This can be seen in the standard result of the PCA transformed data estimated variance,

$$\begin{aligned} S_Z &= \frac{1}{n}Z^TZ = \frac{1}{n}\Omega^TY^T(\Omega^TY^T)^T = \frac{1}{n}\Omega^TY^TY\Omega \\ &= \Omega^TS_Y\Omega = \Omega^T\Omega\Lambda \quad \text{by eigenvector definition} \end{aligned}$$

$$= \quad \Lambda \qquad \text{as } \Omega \text{ orthonormal matrix, i.e. } \Omega^T\Omega = I.$$

Although PCA has been presented so far as a change of basis, it is primary used as a dimension reduction technique. Not all P principal components are usually required to adequately explain the original data. Because the data are summarised by fewer dimensions, PCA can be used to identify the signals of interest in the data visually. To find a (dimension) threshold of principal components, the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_P$ can be examined to find either a steep decline in variance or a threshold of proportionate contribution to the variance. By plotting the transformed data, group differences in the transformed space can become apparent (although the PCA calculation is blinded to group information).

4.1.2 Missing values

Before performing PCA, one further consideration needs to be made. Peak expressions will have non-observed signal (referred to hereafter as missing values) for many spectra. This poses a problem for PCA (and many of the modelling techniques seen herein) as it requires complete data to calculate the eigenvectors. Three common ways of handling missingness are by imputing the missing values as follows.

- Replace the missing values with the mean observed expression of the peak. This is the default method available for the functions in `R` packages `FactoMineR` (Husson et al., 2012) and `missMDA` (Husson and Josse, 2012).
- Replace the missing values with random (Gaussian) observations with a variance estimated by the observed expressions.
- Replace the missing values with a more sophisticated imputation scheme.

Although the first option is simple and will impute constant values within peaks, it works well in practice (de Souto et al., 2015).

Two of the more sophisticated methods of imputation considered here are: k nearest neighbours (kNN) and missMDA imputation. Both methods do not utilise disease membership information, which is an important property required for unbiased discrimination implemented later on. However, by not using class information in imputing missing values, inference on the resulting imputed data is likely to produce conservative and downwards biased estimates of group effects.

The kNN method is implemented in the `R`/`Bioconductor` package `impute` (Hastie et al., 2014) and is described in more detail in Troyanskaya et al. (2001). As the name suggests, missing values are estimated from their k nearest neighbours. For any given missing value in a peak, this method finds the k nearest (peak) neighbours, where the nearest neighbour candidates are limited to peaks with observed values corresponding to the given missing value. The proximity of neighbours are calculated by the Euclidean distance between peaks in $\mathbb{R}^n$ (and missing values contribute the average distance of observed distances). The imputed value is the average of the kNN expressions weighted by the Euclidean distances between the peak with the missing value and the peak neighbour.

The missMDA method is described by Ilin and Raiko (2010) and Audigier et al. (2013) and is similar to the decomposition method of Troyanskaya et al. (2001). Implementation of this method is available in the `R` package `missMDA`. Imputation is performed by assigning initial estimates for missing values, then using a number of principal components deemed significant (by bootstrapping), the values are updated using the expression vectors composed of the principal components. Similarly to kNN imputation, the most closely matched peaks that contain observed values corresponding to the missing value (in the PCA transformed space) determine the imputed value. This process is iteratively applied until a minimum change is achieved.

It should be noted, however, that the less sophisticated scheme of imputation using the mean peak expression provided similar PCA results to those obtained when using kNN imputation and missMDA.

PCA on the GC mice peak expression dataset

Figure 4.1 shows the signal differences in the disease groups of the GC mice data using peak mean values to impute missing values. The pairwise combinations of the first three principal components (contributing to 39.4% of the total variance) are plotted with colour coding of the group (PCA transformed) observations. The two cancer subgroups FF and FFIL6 are the warmer yellow colours and the control groups FFStat3, IL6 and WT are the cooler blue colours. Principal components one and three provide the most differentiation between groups and group means, while principal component two provides very little difference in group expression. However, there is evidence of a chip effect in principal component two. The means for chips 1, 2 and 3 order themselves from left to right within each of the five disease groups in principal component two. Principal component one shows differentiation between the FF group and the remaining groups while principal component three tends to isolate the cancer groups FF and FFIL6 from the remaining groups. The group effect seen in principal components one and three can also be seen to be conflated with a chip effect in principal component three.

Chip information is plotted in Figure 4.1 but is further explored in Figures 4.2 and 4.3. Visualisation of the data provides important exploration of possible batch

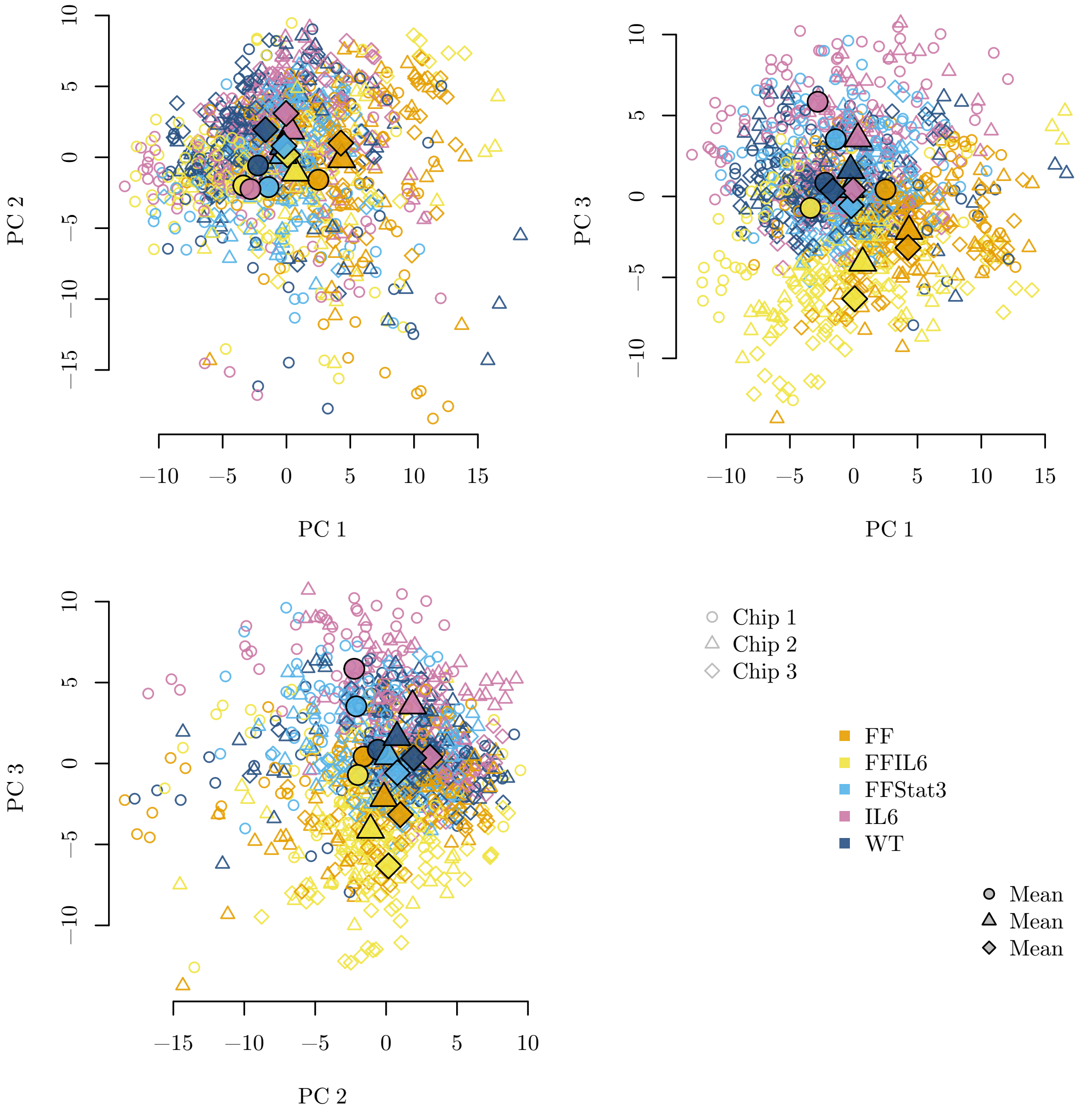


Figure 4.1: PCA plots of GC mice peak expression data with disease classification group ($k = 1, 2, \ldots, 5$) and chip ($j = 1, 2, 3$) labelled. The 1080 PCA points are plotted in a random order irrespective of group membership to avoid a visual bias from plotting points in group order.

effects. In Figure 4.1 it is apparent there is a consistent chip effect within mouse group. For example, in the third principal component, chip 1 has the highest mean down to chip 3 having the lowest within all five mice groups. Chip effects are expected due to the sensitivity in the proteomic MS system to environmental and temporal factors. Figure 4.2 removes the mice disease groupings to focus on the chip effects. The mean values in the first three principal components of the chips are relatively similar when considered within individual principal components. However, in the second principal component especially, there exists some inflated variation of observations on chip 1 where smaller differences are found between chips 2 and 3. The chip effects evident in the principal components will need to be accounted for in the upcoming linear modelling on the expression data (§4.2).

Figure 4.3 utilises the run-order information of the spectra generation on chip 1. In general there does not appear to be an effect of run-order on the peak expression. In the plot of the first and second principal components for chip 1, there are some values which sit away from the central cluster of data but they do not appear to be dominated by either early or late run-order spectra. Figures C.1 and C.2 are the corresponding plots for chips 2 and 3 for reference in Appendix C; however no relationship between run order and the principal components were observed for these chips.

PCA analysis can be used for preliminary biomarker identification. Peak weightings contributing the most (in absolute value) to the principal components exhibiting group differentiation correspond to candidate biomarkers. Table 4.1 provides the peaks with highest weighted contributions to the first three principal components of the GC mice data. Two of the top five weighted contributions of principal component three are identified as potential biomarkers in the linear modelling (§4.2.1) of the GC mice dataset. It is consistent with Figure 4.1 that two biomarker candidates feature as high contributors to principal component three. Principal component three provided the most differentiation between the GC groups.

Table 4.1: Top five peaks contributing to the first three principal components of the GC mice peak expression data.

Principal component	Top 5 peaks (m/z values)				
1	11509	15882	11120	15844	15759
2	5854	2793	3269	8118	3246
3	12281	11855	17458†	8607†	4358

†Peaks that are flagged as peptides of interest in §4.2.1.

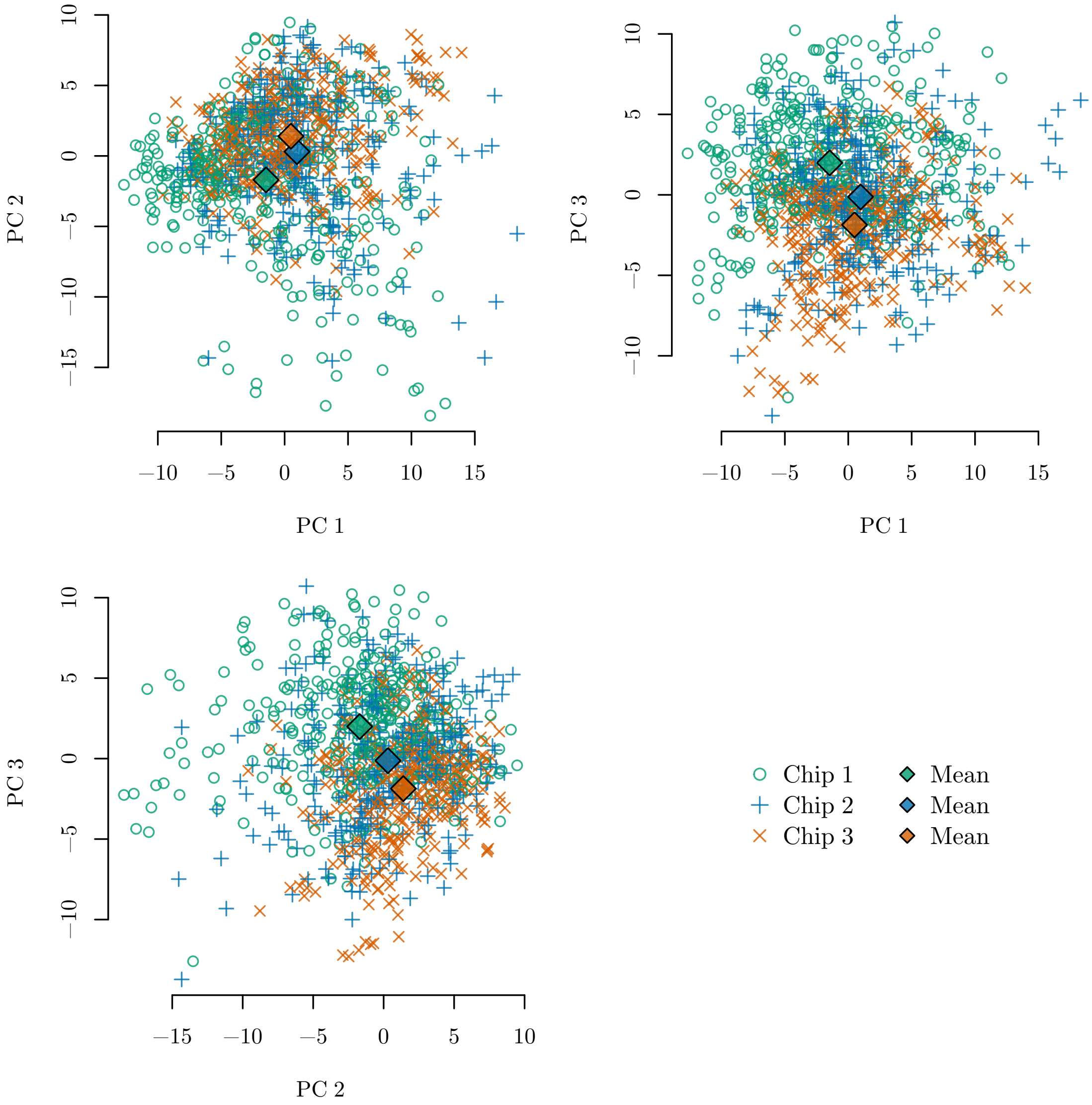


Figure 4.2: PCA plots of GC mice peak expression data with labels for chip only ($j = 1, 2, 3$). The 1080 PCA points are plotted in a random order irrespective of chip number to avoid a visual bias from plotting points in run-order.

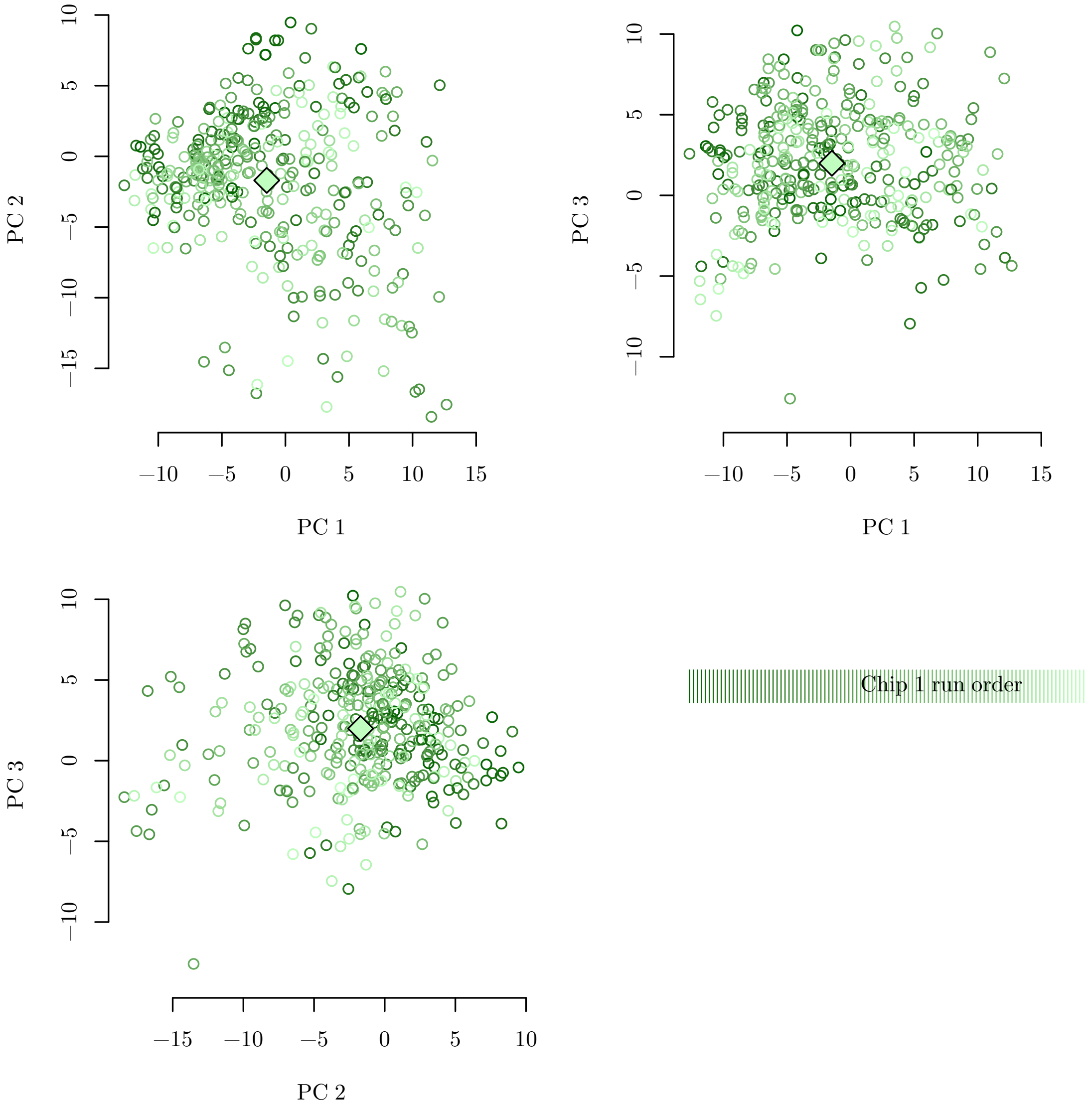


Figure 4.3: PCA plots of GC mice peak expression from chip 1 only with run-order information. The 360 PCA points are plotted in a random order irrespective of chip run-order to avoid a visual bias from plotting points in run-order.

PCA on the other peak expression datasets

The Adam et al. (2002) dataset PCA, shown in Figure 4.4, demonstrates an interesting scatter of PCA points away from the main cluster. Points taking values in the range $[-20, -5]$ in the second principal component mainly correspond to more progressed PC patients (CanB group). The extreme values in the second principal component largely result from the CanB patients having non-missing peak expressions at the low- and high-end of observed peak expression for the highest contributing peaks. There is no observable pattern in the group means in the first three principal components; there is no order in increasing level of disease state, from controls (Cont group) to more developed PC (CanB group), which would suggest a biological gradient between PC disease state and peak expression.

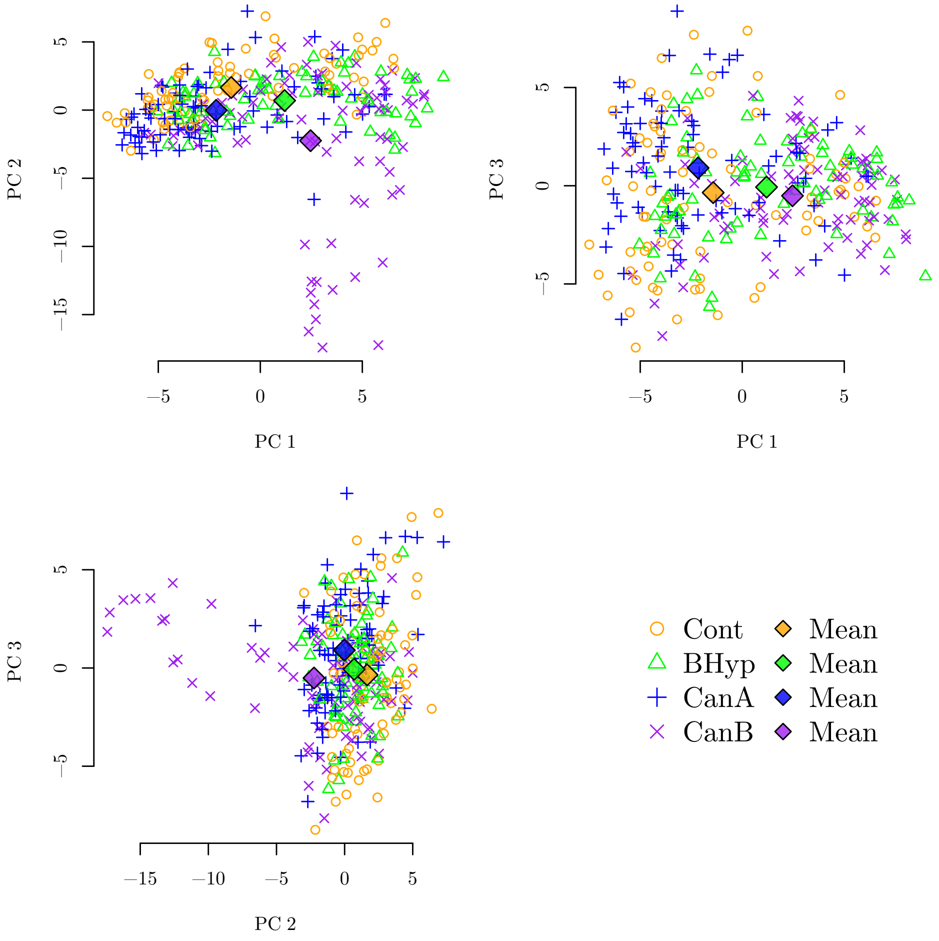


Figure 4.4: PCA plots of peak expression intensities of the Adam et al. (2002) dataset. The 326 PCA points are plotted in a random order irrespective of group membership to avoid a visual bias from plotting points in group order.

The de Noo et al. (2006) dataset shows almost perfect separation of the cancer and control groups for the first two principal components; see Figure 4.5. Such stark group differences are also seen in the differential expression of particular peaks (§4.2.2) and in the low-error discriminant rule to be discussed in Chapter 6.

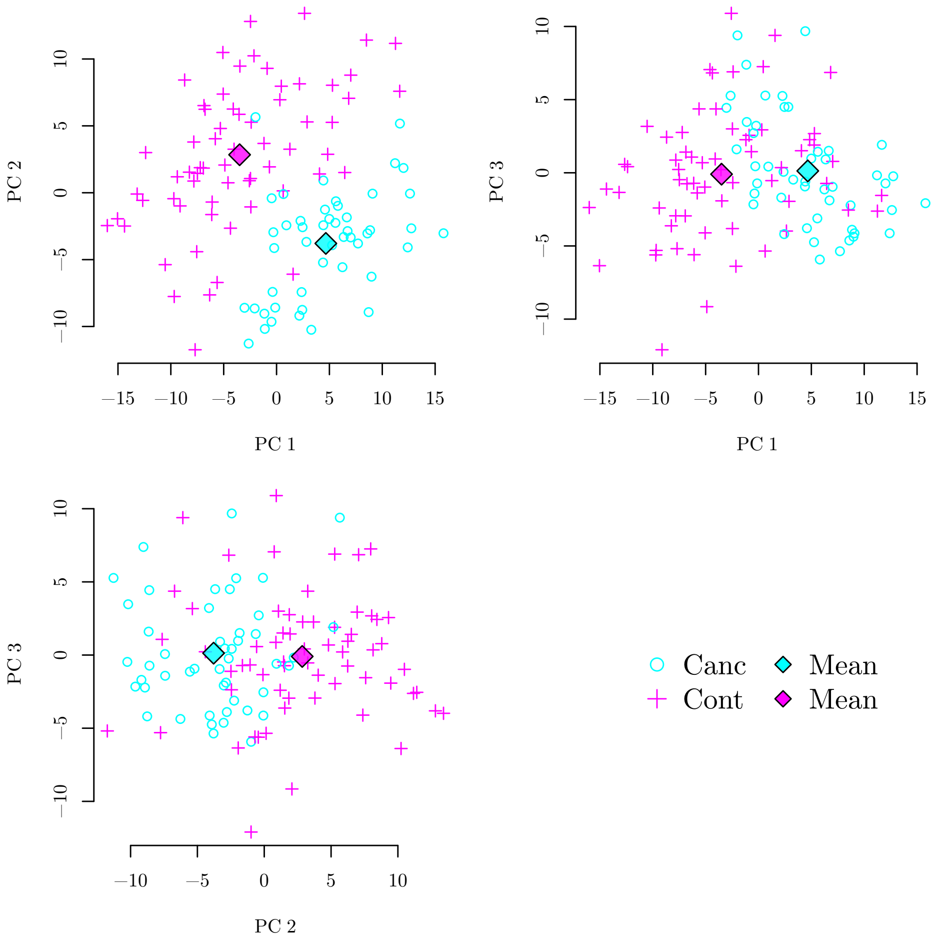


Figure 4.5: PCA plots of peak expression intensities of the de Noo et al. (2006) dataset. The 112 PCA points are plotted in a random order irrespective of group membership to avoid a visual bias from plotting points in group order.

Figures C.3 and C.4 in Appendix C display the results from the PCA of the two asthma peak expression datasets. The asthma1 PCA shows a lack of differentiating peak expression between sex. Similarly, the asthma2 data shows a plot consistent with a lack of class differentiation; this is expected due to the serum sample handling problems discussed previously.

4.2 Using linear models to identify potential biomarkers

Since the expression data, $Y_{(n\times P)}$, are in the form of P peak vectors, separate linear models can be fitted to each n-dimensional vector as the outcome variable (Karpievitch et al., 2012). Peaks in MS experiments are the analogue for genes in microarray experiments analysed by a linear model for each vector of gene expressions (Wolfinger et al., 2001; Smyth, 2005; Rosa et al., 2005). Here, the outcome expression vectors for each peak are regressed on known experimental factors including disease classification, which is the primary variable of interest. From the regression models, the mean effect size associated with group membership and its associated statistical significance can be assessed for each peak.

As discussed previously with MALDI/SELDI-TOF MS data, missingness of peak expressions from the spectra is a hurdle to be overcome in the analysis. This is addressed in §4.3.1 with an assessment of the assumption of *missing at random* (Rubin, 1976) for linear models.

The various expression datasets used in this thesis have different experimental structures. Accordingly, different models are applied.

4.2.1 Linear modelling of the GC mice data

Figure 4.6 presents the GC mice experimental structure, demonstrating the multi-level relationships between the experimental effects. Of most interest is the disease effect on peak expression. However, there are effects that can be attributed to mouse, chip, aliquot, C8 bead treatment as well as random variation. The commonly used linear mixed effects (LME; Laird and Ware, 1982) models are employed to model these data. Although alternate linear modelling methods to account for correlated observations as nuisance variables are available to assess the primary relationship of interest, disease membership and expression, LME models allow for additional investigation into the components of variance.

The expression data

A heatmap of the data for illustrative purposes can be seen in Figure 4.7. The horizontal black lines separate the data into the expression values derived by the three MALDI chips and the five colours on the left denote the genotype associated with the spectrum. Additionally, within each peak, the values have been scaled to $[-1, 1]$ for the heatmap depiction. The analysis is performed on the $\log_2$ expressions

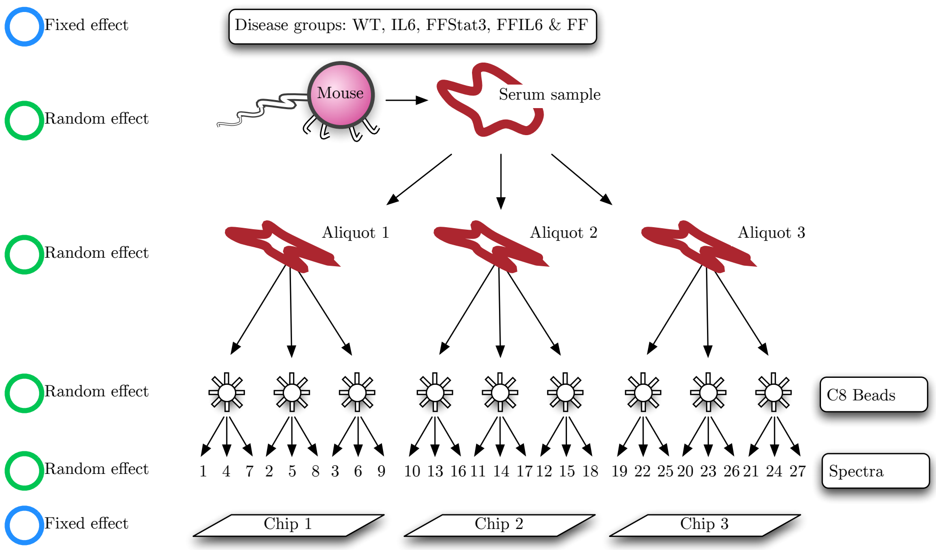


Figure 4.6: Experimental design of GC mice dataset with linear mixed effect model annotation. The numbers 1-27 above the chips are the labels of the 27 spectra produced per mouse.

which will have different mean values for different peaks and are roughly Gaussian in appearance. Note that white cells represent missing values and values scaled to 0 are light pink. Figure 4.7 demonstrates differing expression levels for some peaks in chip 1 (the top third) from the other two chips, as seen in the PCA (§4.1.1). There is also some indication of a chip effect on the occurrence of missing values for some peaks. Some peaks potentially exhibit differential expression between the GC and control groups; however the modelling undertaken in this chapter attempts to account for other sources of variation that might cause such observed differential expression.

The linear model

A linear model of the GC mice data incorporating the experimental structure can be considered a nested, four-level model; a level corresponds to samples taken from a population (Snijders and Bosker, 2012). Mouse is the highest level (level 4), aliquot is the third level (level 3), C8 replicate is the second lowest level (level 2) and technical replicate, or residual error, is the lowest level (level 1). Figure 4.8 demonstrates the differing levels of correlation between the 27 spectra derived from each mouse, a result of the multilevel relationships that exist between spectra. Spectra from the same C8 bead fractionation share more similarity than those from another C8 bead

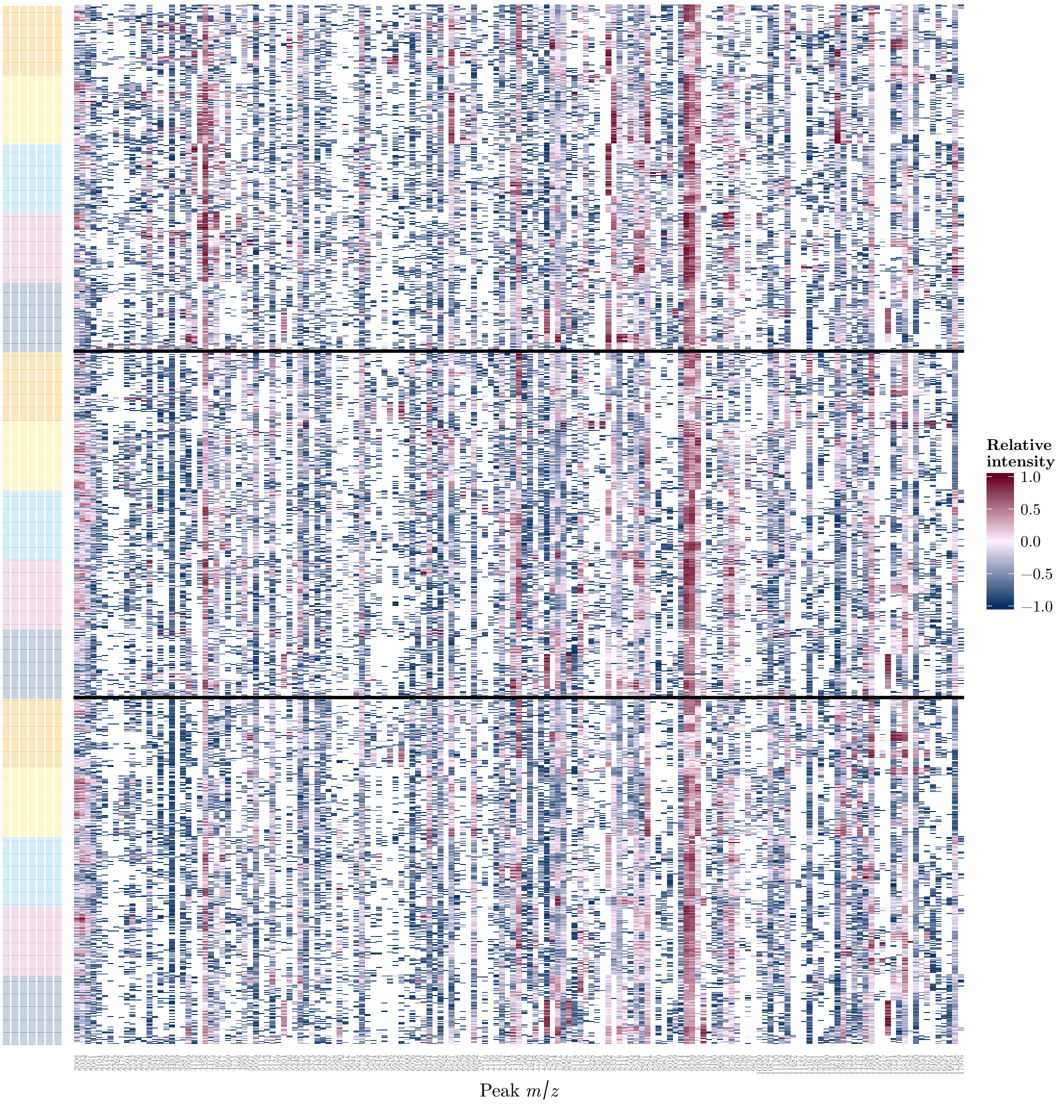


Figure 4.7: GC mice peak expression data as a heatmap; rows correspond to spectra, columns to peaks. Rows are ordered by chip then group membership. The $\log_2$ peak expressions are scaled to $[-1, 1] = [\text{blue}, \text{red}]$ as relative intensity within peak. The row colours orange, yellow, light blue, purple and blue depict the group membership of spectra derived from mice in the FF, FFIL6, FFStat3, IL6 and WT groups, respectively.

fractionation or derived from a different aliquot/MALDI chip. Figure 4.8(a) shows a heatmap of a theoretical correlation structure of a given peak's expression within a mouse for clarity. The darker the cells, the higher the correlation. The empirical correlation matrix in Figure 4.8(b) was constructed by calculating a matrix of all pairwise correlations of peak expressions for the 27 replicates per mouse, then taking the average correlation matrix over the 40 mice. The empirical correlation matrix was remarkably similar to the expected correlation matrix that was hypothesised from the nested experimental structure. Expressions are assumed independent between mice.

The definition of what constitutes a random effect is varied (Gelman, 2005). Here random effects are considered to be samples from a broader population (Snijders and Bosker, 2012). Therefore, mouse, aliquot, C8 treatment and technical residual error are considered random effects as labelled previously in Figure 4.6. Each mouse within disease classification is a sample from a population of mice. Each aliquot can be considered a sample from a population of possible serum samples from each mouse. C8 beads are applied as a single use solution to affinity capture proteins for the MS, and nine different C8 bead solutions were used for each mouse. Thus, C8 bead treatment can be considered a sample from a population of C8 bead solutions. Finally, the technical replicates can be considered as samples from a population of all possible technical replicates.

The disease group and chip effect for each peak are assumed fixed effects (Figure 4.6). Chip and aliquot are confounded as aliquot shares a one-to-one relationship with chip. Chip and aliquot effects are nominally differentiated as chip is treated as fixed and aliquot is treated as random.

For each peak $p = 1, 2, \ldots, 159$ separately, the expression value, $Y_{pijk\ell}$, for mouse i, aliquot j and C8 bead treatment k and replicate ℓ can be modelled,

$$Y_{pijk\ell} = \underbrace{\mu_p + \gamma_{pj} + \eta_{pg_i}}_{\text{fixed effects}} + \underbrace{\xi_{pi} + \varphi_{pij} + \psi_{pijk} + \epsilon_{pijk\ell}}_{\text{random effects}}, \tag{4.1}$$

where

μ_p is the mean peak expression fixed effect;

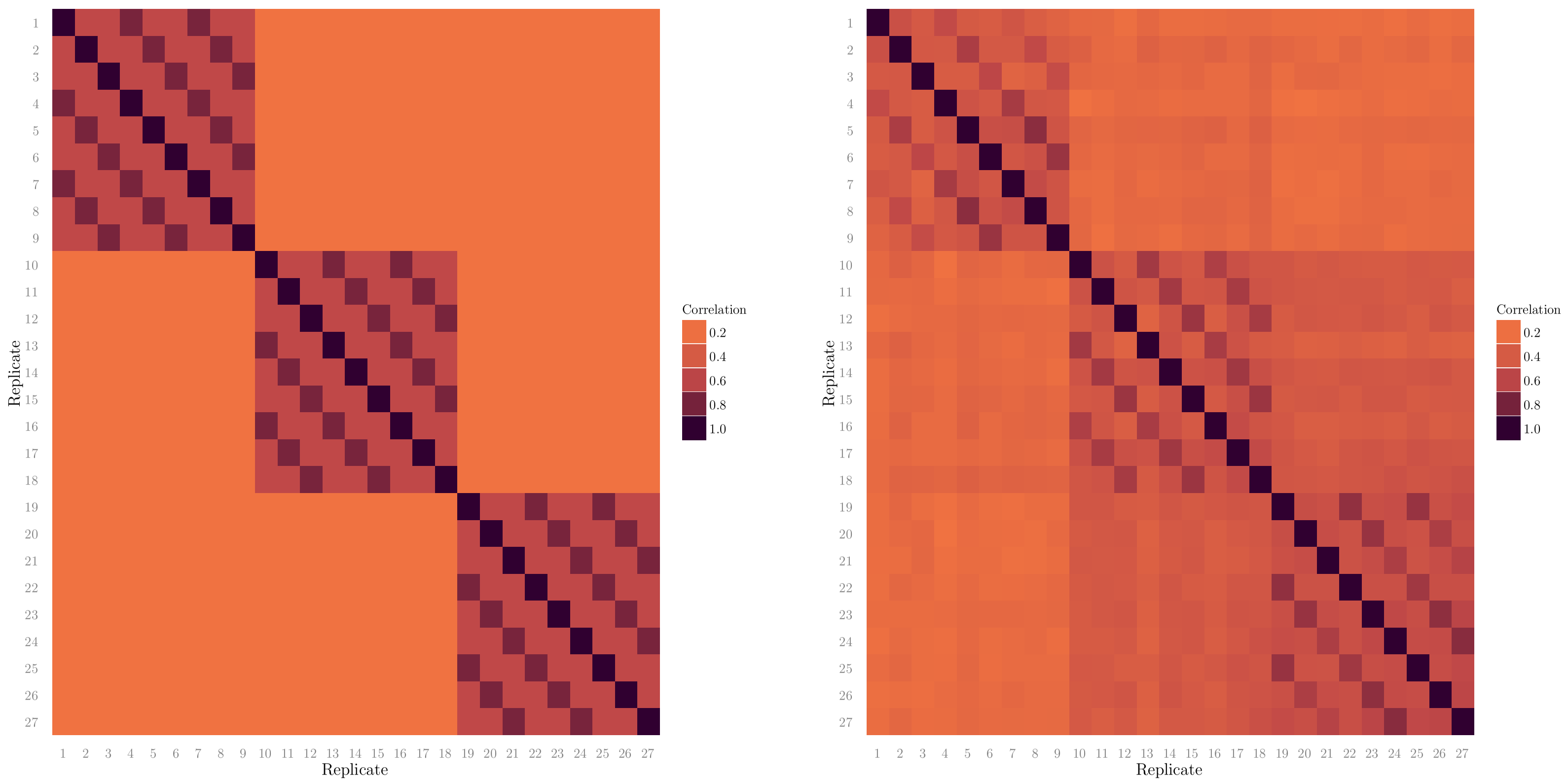
γ_{pj} is the MALDI chip $j = 1, 2, 3$ fixed effect;

η_{pg_i} is the disease group $g_i \in \{\text{FF}, \text{FFIL6}, \text{FFStat3}, \text{IL6}, \text{WT}\}$ fixed effect for corresponding to mouse i;

$\xi_{pi} \sim N(0, \tau_{p3}^2)$ is the level 4 random effect for mouse $i = 1, 2, \ldots, 40$;

$\varphi_{pij} \sim N(0, \tau_{p2}^2)$ is the level 3 random effect for aliquot $j = 1, 2, 3$, in mouse i;

$\psi_{pijk} \sim N(0, \tau_{p1}^2)$ is the level 2 random effect for C8 treatment $k = 1, 2, 3$, in aliquot j and mouse i; and,



(a) Theoretical

(b) Empirical

Figure 4.8: Theoretical and empirical pairwise correlation structure of peak expression for each set of 27 spectra for each mouse in the GC mice dataset. The correlation can be seen in descending strength in colour change from dark to light.

$\epsilon_{pijk\ell} \sim N(0, \sigma_p^2)$ is the level 1 residual error for replicates $\ell = 1, 2, 3$, in C8 treatment k, aliquot j and mouse i.

There are $n = \sum_{i=1}^{40} \sum_{j=1}^{3} \sum_{k=1}^{3} \sum_{\ell=1}^{3} n_{ijk\ell} = 1080$ expression values $y_{pijk\ell}$ for a given peak p, when all peak expression values are observed.

The random effects ξ_{pi}, φ_{pij}, ψ_{pijk} and $\epsilon_{pijk\ell}$, are assumed to be independently normally distributed within and between random effects. The model parameters τ_{p3}^2, τ_{p2}^2, τ_{p1}^2 and σ_p^2 are referred to as the *variance components* and provide insight into the variability of expression values imposed by each level of the experimental design.

The chip and disease group fixed effects are not individually identifiable so they are parameterised to be contrasts to a designated level of the variable. Chip 1 was chosen as the reference group for chip and WT mice was the reference group for disease group.

Matrix representation of the linear model

The LME model in Equation (4.1) can be written in vector notation as

$$\mathbf{Y}_p = \mathbf{X}\boldsymbol{\beta}_p + \mathbf{Z}\mathbf{B}_p + \boldsymbol{\epsilon}_p, \tag{4.2}$$

where $\mathbf{Y}_p$ is the n-dimensional outcome vector of the p^{th} peak's expressions, $\mathbf{X}$ is an $n \times d$ design matrix of indicators corresponding to the fixed effects, $\boldsymbol{\beta}_p = (\mu_p, \gamma_{p1}, \gamma_{p2}, \eta_{p1}, \ldots, \eta_{p4})^T$ is a d-dimensional fixed effects model parameter vector, $\mathbf{Z}$ is an $n \times q$ design matrix of indicators corresponding to the random effect parameters, $\mathbf{B}_p = (\xi_{p1}, \ldots, \xi_{p40}, \varphi_{p1,1}, \ldots, \varphi_{p40,3}, \psi_{p1,1,1}, \ldots, \psi_{p40,3,3})^T$ is a q-dimensional random effect model parameter vector and $\boldsymbol{\epsilon}_p$ is an n-dimensional vector of errors (McLean et al., 1991; Fox, 2002; Pinheiro and Bates, 2009). The matrices $\mathbf{X}$ and $\mathbf{Z}$ are constant over all P peaks because the experimental design is imposed on all peaks simultaneously. The number of columns in $\mathbf{X}$ and $\mathbf{Z}$ are $d = 1 + 2 + 4 = 7$ and $q = 40 + 40 \times 3 + 40 \times 3 \times 3 = 520$, respectively.

The distribution of $\mathbf{Y}_p$, conditional on observed random effects $\mathbf{b}_p$, is multivariate Gaussian (Bates et al., 2014a),

$$\mathbf{Y}_p | \mathbf{B}_p = \mathbf{b}_p \sim \mathcal{N}_n \left(\mathbf{X}\boldsymbol{\beta}_p + \mathbf{Z}\mathbf{b}_p, \sigma_p^2 \mathbf{I} \right).$$

The vector $\mathbf{b}_p$ is an observation of the random variable $\mathbf{B}_p$ with distribution,

$$\mathbf{B}_p \sim \mathcal{N}_q \left(\mathbf{0}, \boldsymbol{\Sigma}_p \right),$$

where $\boldsymbol{\Sigma}_p$ depends on $\boldsymbol{\theta}_p = \left\{ \tau_{p3}^2, \tau_{p2}^2, \tau_{p1}^2 \right\}$. The variance-covariance matrix $\boldsymbol{\Sigma}_p$ here is a diagonal matrix containing components $\tau_{p3}^2, \tau_{p2}^2, \tau_{p1}^2$ corresponding to the random effects $\xi_{pi}, \varphi_{pij}, \psi_{pijk}$, respectively, in $\mathbf{B}_p$.

Solutions to a linear mixed effects model and the `lme4` package

The LME formulation requires estimates of the model parameters $\boldsymbol{\theta}_p$, $\boldsymbol{\beta}_p$, $\mathbf{b}_p$ and σ_p^2. The estimates, $\hat{\boldsymbol{\beta}}_p$ and $\hat{\mathbf{b}}_p$, are obtained via an iterative algorithm as no analytical solution exists if the variance components are not known (Laird and Ware, 1982).

Estimates are solutions to the penalised least-squares (PLS) formulation; that is, values that minimise the conditional residuals with an additional penalty term for the magnitude of the random effect estimates.

From the PLS formulation, $\hat{\boldsymbol{\beta}}_p$ and $\hat{\mathbf{b}}_p$, are values satisfying the following $d+q$ simultaneous equations (Robinson, 1991; Venables and Ripley, 2002), referred to as the *normal equations* (West et al., 2007; Bates et al., 2014a),

$$\begin{bmatrix} \frac{1}{\sigma_p^2}\mathbf{Z}^T\mathbf{Z} + \sigma_p^2\boldsymbol{\Sigma}^{-1} & \frac{1}{\sigma_p^2}\mathbf{Z}^T\mathbf{X} \\ \frac{1}{\sigma_p^2}\mathbf{X}^T\mathbf{Z} & \frac{1}{\sigma_p^2}\mathbf{X}^T\mathbf{X} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}}_p \\ \hat{\boldsymbol{\beta}}_p \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_p^2}\mathbf{Z}^T\mathbf{y}_p \\ \frac{1}{\sigma_p^2}\mathbf{X}^T\mathbf{y}_p \end{bmatrix}. \tag{4.3}$$

However, Equation (4.3) relies on σ_p^2 and $\boldsymbol{\theta}_p$ being known or estimated.

The `lme4` package (Bates et al., 2014b) redefines the random variable, $\mathbf{B}_p$, to simplify the distribution of the random effects to be estimated. The random effects variance-covariance matrix, $\boldsymbol{\Sigma}_p$, can be decomposed as $\sigma_p^2\boldsymbol{\Lambda}_\theta\boldsymbol{\Lambda}_\theta^T$ where $\boldsymbol{\Lambda}_\theta$ is called the 'relative covariance factor' (Bates et al., 2014a). Here, $\boldsymbol{\Lambda}_\theta = \boldsymbol{\Lambda}_\theta^T$ is a diagonal matrix with elements τ_{p3}/σ_p, τ_{p2}/σ_p and τ_{p1}/σ_p. The relative covariance factor is used to redefine the unobserved random effects $\mathbf{B}_p = \boldsymbol{\Lambda}_\theta\mathbf{U}_p$, where $\mathbf{U}_p \sim \mathcal{N}_q\left(\mathbf{0}, \sigma_p^2\mathbf{I}\right)$ are called 'spherical random effects' (Bates et al., 2014a). The conditional distribution of $\mathbf{Y}_p$ can therefore be re-expressed,

$$\mathbf{Y}_p|\mathbf{U}_p = \mathbf{u}_p \sim \mathcal{N}_n\left(\mathbf{X}\boldsymbol{\beta}_p + \mathbf{Z}\boldsymbol{\Lambda}_\theta\mathbf{u}_p, \sigma_p^2\mathbf{I}\right).$$

With the transformation of the random effect variable, the normal equations of Equation (4.3) now simplify to,

$$\begin{bmatrix} \boldsymbol{\Lambda}_\theta^T\mathbf{Z}^T\mathbf{Z}\boldsymbol{\Lambda}_\theta + \mathbf{I} & \boldsymbol{\Lambda}_\theta^T\mathbf{Z}^T\mathbf{X} \\ \mathbf{X}^T\mathbf{Z}\boldsymbol{\Lambda}_\theta & \mathbf{X}^T\mathbf{X} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_p \\ \hat{\boldsymbol{\beta}}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}_\theta^T\mathbf{Z}^T\mathbf{y}_p \\ \mathbf{X}^T\mathbf{y}_p \end{bmatrix}, \tag{4.4}$$

as $\mathbf{u}_p$ is estimated in lieu of $\mathbf{b}_p$. The normal equations no longer rely on the inverse of Σ (or Λ_θ). This avoids computational singularities associated with finding matrix inverses as well as the computational burden involved.

Speed improvements in estimation can be achieved utilising the Cholesky decomposition, as implemented in `lme4`, of the normal equations. The normal equations in

Equation (4.4) can be re-defined,

$$\begin{bmatrix} \mathbf{L}_\theta^T & \mathbf{R}_{ZX} \\ \mathbf{0} & \mathbf{R}_X \end{bmatrix}^T \begin{bmatrix} \mathbf{L}_\theta^T & \mathbf{R}_{ZX} \\ \mathbf{0} & \mathbf{R}_X \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_p \\ \hat{\boldsymbol{\beta}}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}_\theta^T \mathbf{Z}^T \mathbf{y}_p \\ \mathbf{X}^T \mathbf{y}_p \end{bmatrix}, \tag{4.5}$$

where $\mathbf{L}_\theta$ is a lower triangular matrix and $\mathbf{R}_X$ is an upper triangular matrix from the Cholesky decomposition (Bates et al., 2014a).

Estimation of $\boldsymbol{\theta}_p$ is made with use of its profiled log-likelihood (Bates, 2011); the value of $\hat{\boldsymbol{\theta}}_p$ that minimises the restricted maximum likelihood (REML, unbiased opposed to standard ML) criterion,

$$-2\mathcal{L}_R\left(\boldsymbol{\theta}_p|\mathbf{y}_p\right) = \log\left\{|\mathbf{L}_\theta|^2|\mathbf{R}_X|^2\right\} + (n-d)\left[1 + \log\left(\frac{2\pi r^2(\boldsymbol{\theta}_p)}{n-d}\right)\right],$$

where $r^2(\boldsymbol{\theta}_p)$ is the residual sum of squares (conditioned on the random effects, penalised by $||\mathbf{u}_p||^2$) for estimates of $\hat{\mathbf{u}}_p$ and $\hat{\boldsymbol{\beta}}_p$.

The estimates of $\boldsymbol{\beta}_p$ and $\mathbf{u}_p$ depend on a known value for $\boldsymbol{\theta}_p$ in the normal equations, thus an iterative process successively updating values determined from the non-linear minimisation the REML criterion and the solutions to the normal equations is required. The final estimates of $\boldsymbol{\beta}_p$ and $\mathbf{u}_p$ are made using the final estimate of $\boldsymbol{\theta}_p$ that achieves convergence in the REML criterion. The REML estimate for σ_p^2 is simply $\frac{r^2(\hat{\boldsymbol{\theta}}_p)}{n-d}$ (Bates et al., 2014a).

The package `lme4` uses 'general purpose' non-linear optimisers that do not use the REML criterion gradients with respect to $\boldsymbol{\theta}_p$. The suggested methods are the Nelder-Mead simplex method (Nelder and Mead, 1965) or the more recently developed BOBYQA method (Powell, 2009).

Missingness and variance components

The model in Equation (4.1) assumes all $n = 1080$ peak expression values for peak p are observed. However, the GC mice expression data have missing values in the outcome vector. Missing values require that the entries in the outcome vector and the corresponding rows of the design matrices are removed. If the removal of particular rows from the design matrices create columns of zeros, these columns and corresponding elements in $\boldsymbol{\beta}_p$ or $\mathbf{b}_p$ require removal from the model formulation.

So that the results for each peak are directly comparable, a constant model structure was sought to model each peak. For peaks with a high proportion of missingness, the model proposed in Equation (4.1) might be unnecessarily complex and result in overfitting. As such, additional random effect structures were considered. Figure C.5 in Appendix C shows the different random effect models considered with the model fit metrics of the Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) and residual variance for each model corresponding to the P peaks. Minimised BIC and AIC indicate a preferred model, they are functions of the negative log-likelihood (Schwarz, 1978). It was apparent the peaks with increased missingness had lower BIC values for simpler models, as is the tendency with the BIC metric to favour fewer parameters for smaller samples.

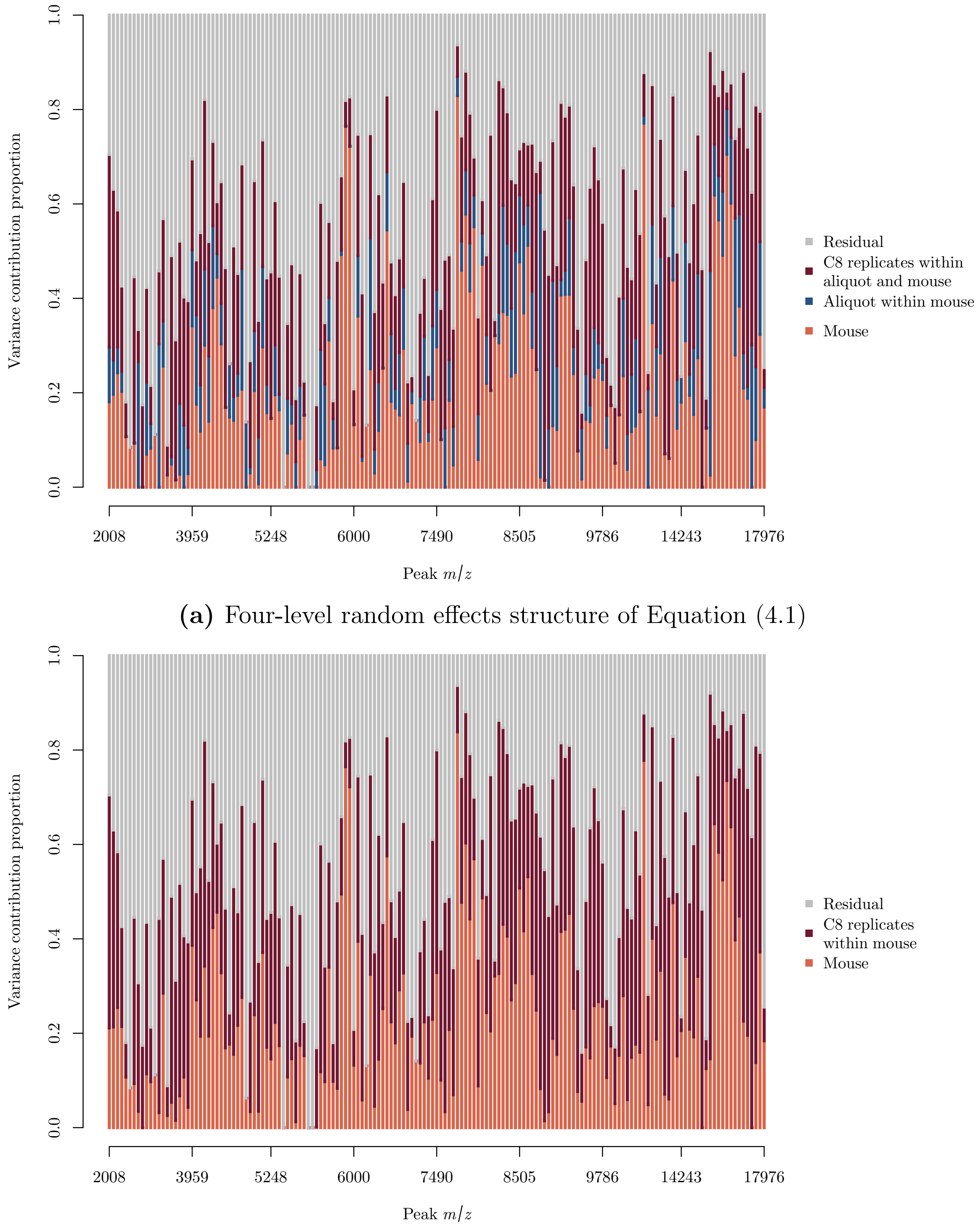
The four-level model of Equation (4.1) had the minimum BIC and AIC of all the random effect structures considered in the majority of cases of the $P = 159$ linear models. However, 28% of these models were 'degenerate' (Bates, 2010); models where the (RE)ML estimate of at least one of the variance components is 0. While still a valid model (and valid parameter estimates); a consistent, parsimonious model structure was sought. From the AIC and BIC plots (Figure C.5 in Appendix C), the subsequently preferred model was a reduced three-level model. It should be noted, this model minimised the BIC and AIC for the majority of low sample size peaks. The three-level model took a similar form to Equation (4.1), however no random effect term for aliquot (φ_{pij}) was included. The three-level model takes the form,

$$Y_{pik\ell} = \underbrace{\mu_p + \gamma_{pj} + \eta_{pg_i}}_{\text{fixed effects}} + \underbrace{\xi_{pi} + \psi_{pik} + \epsilon_{pik\ell}}_{\text{random effects}}, \tag{4.6}$$

where the parameters are previously defined, however $\psi_{pik} \sim N(0, \tau_{p1}^2)$ for C8 treatments, $k = 1, 2, \ldots, 9$, in mouse i and ξ_{pi} is a third level mouse random effect. Figure C.6 in Appendix C presents a schematic of this formulation in matrix form.

Using the reduced model in Equation (4.6), only 7% of the 159 models were degenerate. We considered this satisfactory as the remaining degenerate models were on peaks with high missingness, with some unable to estimate variation beyond residual variance. It should be noted, no material difference was seen in the fixed effect parameter estimates between the four- and three-level formulations to affect the inference made in the next section.

The variance component contributions can be seen in Figure 4.9 for all $P = 159$ peaks for models in Equations (4.1) and (4.6). The residual variance is a large proportion of the variability observed for both models. The large residual variance is likely to be a result of an under-specified model because of unknown covariates. For the four-level model, the smallest contributions to the variance were generally the third level components of aliquot, suggesting the partitioned serum samples are largely homogeneous. Such small aliquot variability is also consistent with the three-level model selection, removing aliquot as the least required variance component in the four-level formulation. The estimated variance components for the four-level and three-level models show largely unchanged estimates of the residual and mouse variance, with the aliquot variance seen in the four-level model absorbed into the C8 variance of the three-level model.



(a) Four-level random effects structure of Equation (4.1)

(b) Three-level random effects structure of Equation (4.6)

Figure 4.9: Proportional random effect contributions to the variance for each of the 159 LME peak models for (a) the four-level model of Equation (4.1) and (b) the three-level model of Equation (4.6).

Fixed effects for identifying potential biomarkers

The peaks deemed to have the highest biomarker potential were peaks that exhibited both significant differential mean expression between disease groups and met a minimum threshold of mean expression fold change between the cancer and control groups. Significant differential expression, for each peak $p = 1, 2, \ldots, P$, was tested using an F-test assessing the null hypothesis, H_0: all $\eta_{p,g}$ the same, where $\eta_{p,g}$ represents the disease group fixed effect of Equation (4.6) for $g =$ WT, IL6, FFStat3, FFIL6, FF. The corresponding p-values were adjusted to maintain a false discovery rate (FDR) at 0.05 using the Benjamini & Hochberg method (Benjamini and Hochberg, 1995). Post-hoc fold change was estimated using model estimated disease group fixed effects. The mean ($\log_2$) fold change of one and a half is a standard clinical threshold (Griffin et al., 2003; Old et al., 2005) and of most interest is the difference between the cancer and non-cancer phenotypes. The LME model for peak p has a fixed effects vector containing the estimates of,

$$\begin{aligned}
\eta_{p,FF} &- \eta_{p,WT}, \\
\eta_{p,FFIL6} &- \eta_{p,WT}, \\
\eta_{p,FFStat3} &- \eta_{p,WT}, \text{ and} \\
\eta_{p,IL6} &- \eta_{p,WT}.
\end{aligned}$$

The mean fold change for peak p, FC_p, between GC and control mice on the $\log_2$-scale can therefore be represented as

$$\begin{aligned}
\text{FC}_p &= \frac{1}{2}\left((\eta_{p,FF} - \eta_{p,WT}) + (\eta_{p,FFIL6} - \eta_{p,WT})\right) \\
&\quad -\frac{1}{3}\left((\eta_{p,FFStat3} - \eta_{p,WT}) + (\eta_{p,IL6} - \eta_{p,WT})\right) \\
&= \frac{1}{2}\left(\eta_{p,FF} + \eta_{p,FFIL6}\right) - \frac{1}{3}\left(\eta_{p,FFStat3} + \eta_{p,IL6} + \eta_{p,WT}\right).
\end{aligned}$$

The volcano plot (Cui and Churchill, 2003) in Figure 4.10 presents the relationship between peak significance and fold change. The estimate of FC_p can be seen on the x-axis for each peak expression model and the Benjamini & Hochberg adjusted p-values corresponding to the null hypothesis, H_0 : all $\eta_{p,i}$ the same, can be seen on the $\log_{10}$-scale on the y-axis. Table C.1 in Appendix C provides a summary of the information depicted in the volcano plot for reference.

The peaks of most interest are those exhibiting a biological gradient between the five mouse genotypes and peak expression. That is, peaks where the mean expression of the groups are ordered by disease severity. A peak where the mean expression for the two GC cancer groups are similar, that are in turn, different from the three similarly mean expressed non-cancer groups, is also of interest. From Figure 4.11, the peaks

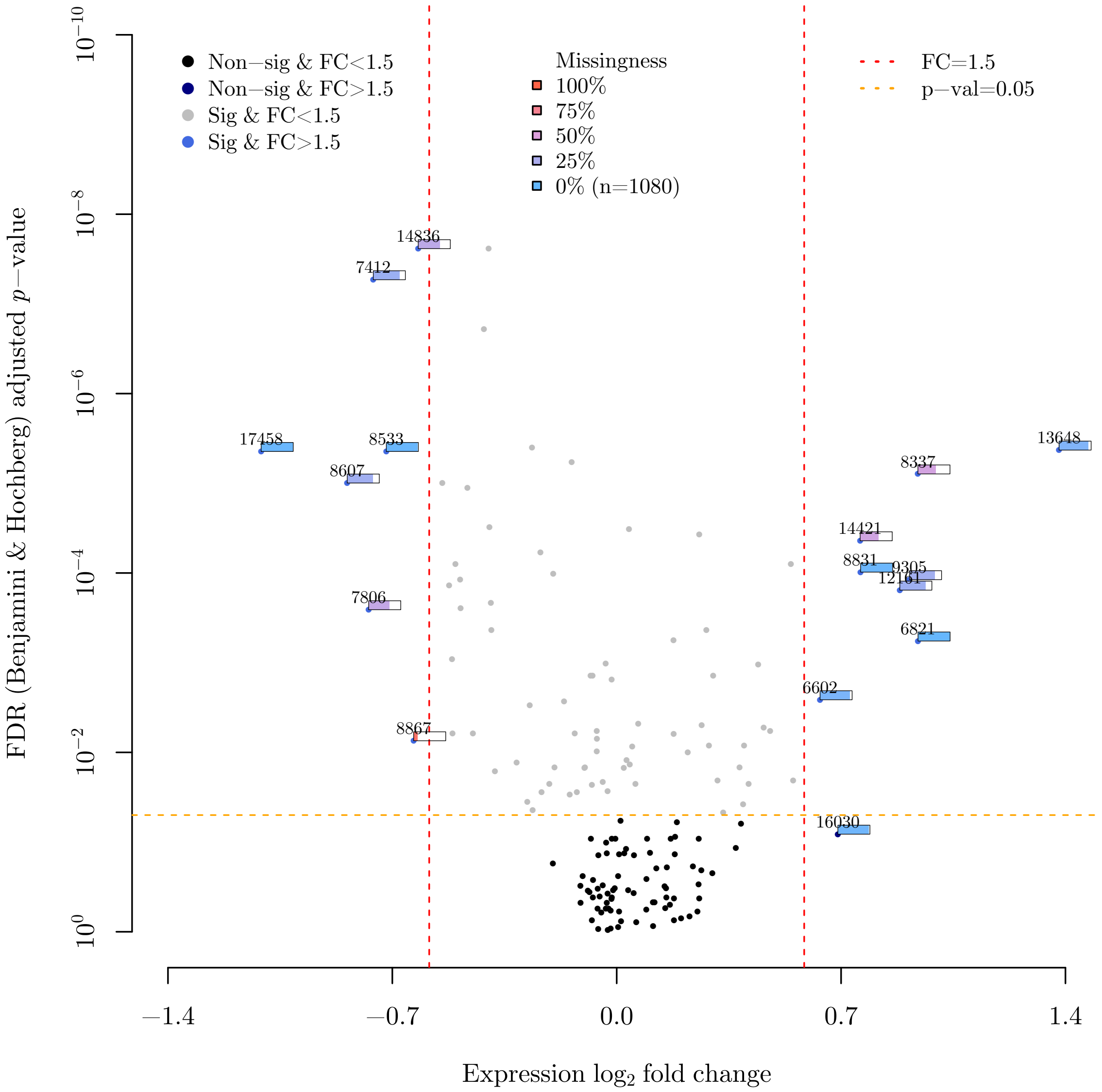


Figure 4.10: Volcano plot of phenotype group differences for each peak of the GC mice expression data; adjusted p-value vs. fold change on the $\log_2$-scale. Missingness observed for peaks with fold-change greater than one and a half is indicated by rectangle fill adjacent to point.

showing biomarker potential are at 6602, 6821, 8607, 13648, 14421 and 17458m/z. Peaks at 6821 and 13648m/z are most likely to be the same underlying protein for which 6821 is a doubly-charged version of a 13648Da peptide as $(13648 + 2\text{H}^+)/2 \approx 6821m/z$. The expressions of these two peaks are highly correlated ($\rho = 0.83$) and show very similar fold change. While the peak at 7806m/z does not show differential expression between the GC groups and the FFStat3 and IL6 groups, it does exhibit a strong mean up-regulated expression for the WT group compared to all other groups.

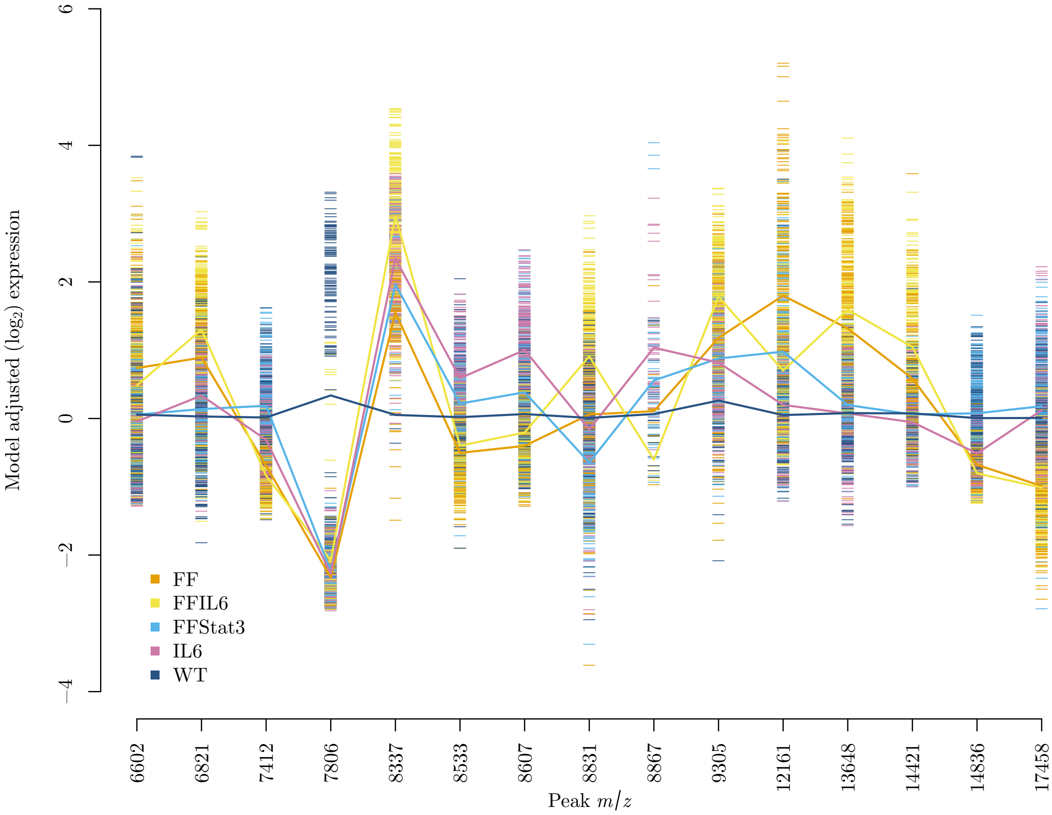


Figure 4.11: Parallel plot of the GC mice peak expression data for peaks identified in Figure 4.10. Model-estimated intercept and chip effects for each peak have been removed for clarity.

4.2.2 Regression on the other peak expression data

Ordinary linear least-squares regression (OLS; Fox, 2002) was fitted to the Adam et al. (2002) and de Noo et al. (2006) peak expression data separately for each peak. As each spectrum is derived from a different patient in these datasets, no experimental random effect structure was known. Missing data are expected to be present in the outcome variable and cannot be used. The model is fitted by removing those missing observations. Intuitively, should the missingness be occurring systematically, the estimated regression coefficients will be biased. Residual plots for the P OLS models for each dataset were used to assess the assumptions of the models, the assumptions seemed reasonable.

Figures 4.12 and 4.13 show the volcano plots corresponding model estimates of disease group peak expression difference and significance. For the de Noo et al. (2006) dataset, the fold differences were calculated between the two experimental groups (cancer or control); the p-value relates to the corresponding t-test. For the Adam et al. (2002) dataset, the fold difference was calculated in a similar fashion to that for the GC mice dataset, namely the control and hyperplasia groups versus cancer A and B groups. The p-value relates to the F-test of the aforementioned disease groups' mean expressions all being equal. All p-values were controlled at a FDR of 0.05 using a Benjamini & Hochberg multiple comparison adjustment.

Tables of the identified peaks from the regression analysis seen in Figures 4.12 and 4.13 are available in Appendix C in Tables C.3 and C.4, with corresponding fold-change and significance values for reference.

A fold change of two on the $\log_2$-scale was used for the de Noo et al. (2006) dataset to limit the number of peaks meeting the statistical significance and minimum fold change criteria. The peaks identified were consistent with previously published results, such as Alexandrov et al. (2009). Seven of the eight peaks identified in Alexandrov et al. (2009) were identified in this analysis, namely the peaks at 1208, 1265, 1352, 1692, 1780, 1867 and $2024m/z$.[1] This analysis however, highlighted previously unidentified potential biomarkers. To add clinical relevance, this analysis provides fold change estimates that were absent in the analysis conducted by Alexandrov et al. (2009). Note the m/z-values presented here may vary by one or two m/z units from other results as an artefact of the peak alignment process in estimating the peak's true location on the m/z-axis. Something unusual about this expression data is that of the 18 identified peaks, all of the up-regulated peak expressions for the cancer group are the highest m/z-values and the down-regulated expressions for the cancer group are the lowest m/z-values, with the exception of two down-regulated peaks for high m/z values. Some p-values corresponding to expression difference hypothesis tests reached the machine precision available in R and can be seen as a ceiling in Figure 4.12 (as well as in Figure 4.13 for the Adam et al. (2002) dataset).

[1] The peak at $1467m/z$ not identified here had neither a significant p-value or a fold change two or greater (adjusted p-value=0.757, FC=0.1). However, this peak had a large proportion of missingness (0.61) for which all but two of the control group expression values were missing. A Fisher's Exact test of counts for observed or absent expression at peak $1467m/z$ between the control and cancer groups was highly significant (p-value=3.6×10^{-12}).

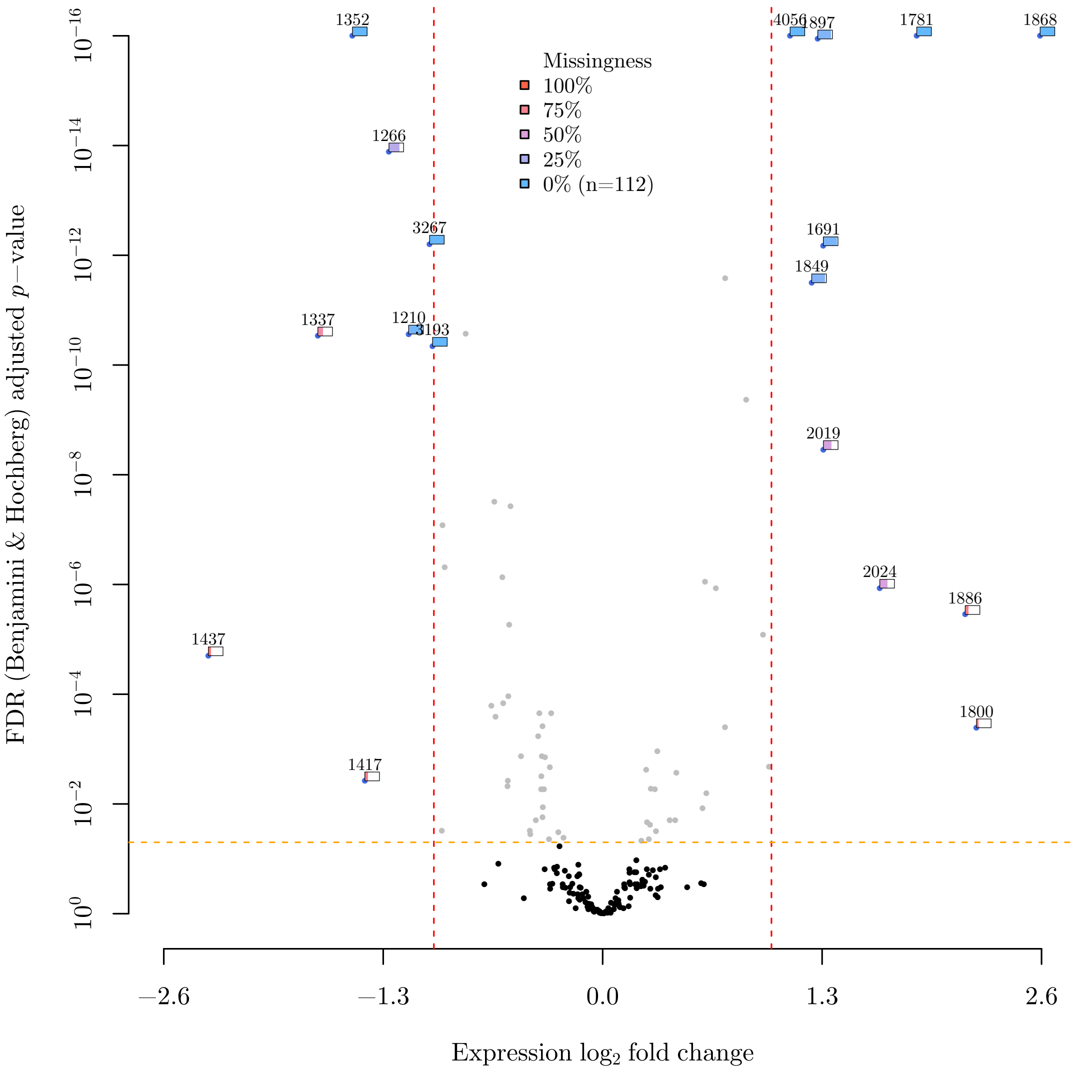


Figure 4.12: Volcano plot for the de Noo et al. (2006) dataset for group differences peak expressions; adjusted p-value vs. fold change on the $\log_2$-scale. Missingness observed for each peak is indicated by rectangle fill adjacent to point.

While the Adam et al. (2002) paper focussed on the discrimination of the disease groups as opposed to statistical identification of biomarkers based on expression difference between groups, only two peaks (8141 and 9149m/z, labelled here as 8142 and 9150m/z, respectively) were re-identified by this analysis. The seven other peaks used in the classification tree of the Adam et al. (2002) paper did not reach the statistical significance and the fold-change threshold. A total of 18 peaks were identified as potential biomarkers in this analysis and can be seen in Figure 4.13 and are summarised in Table C.4 for reference. However, these results should be treated

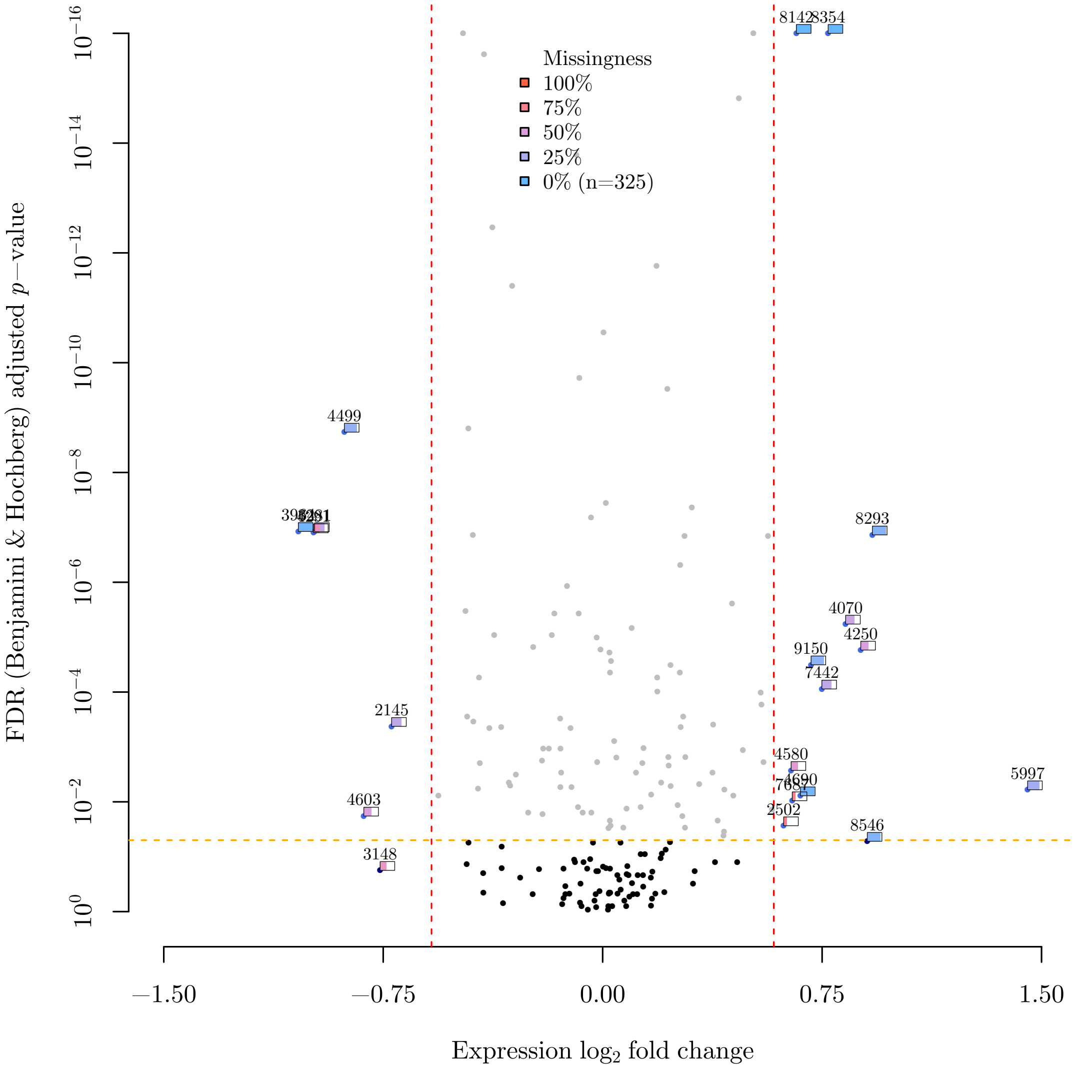
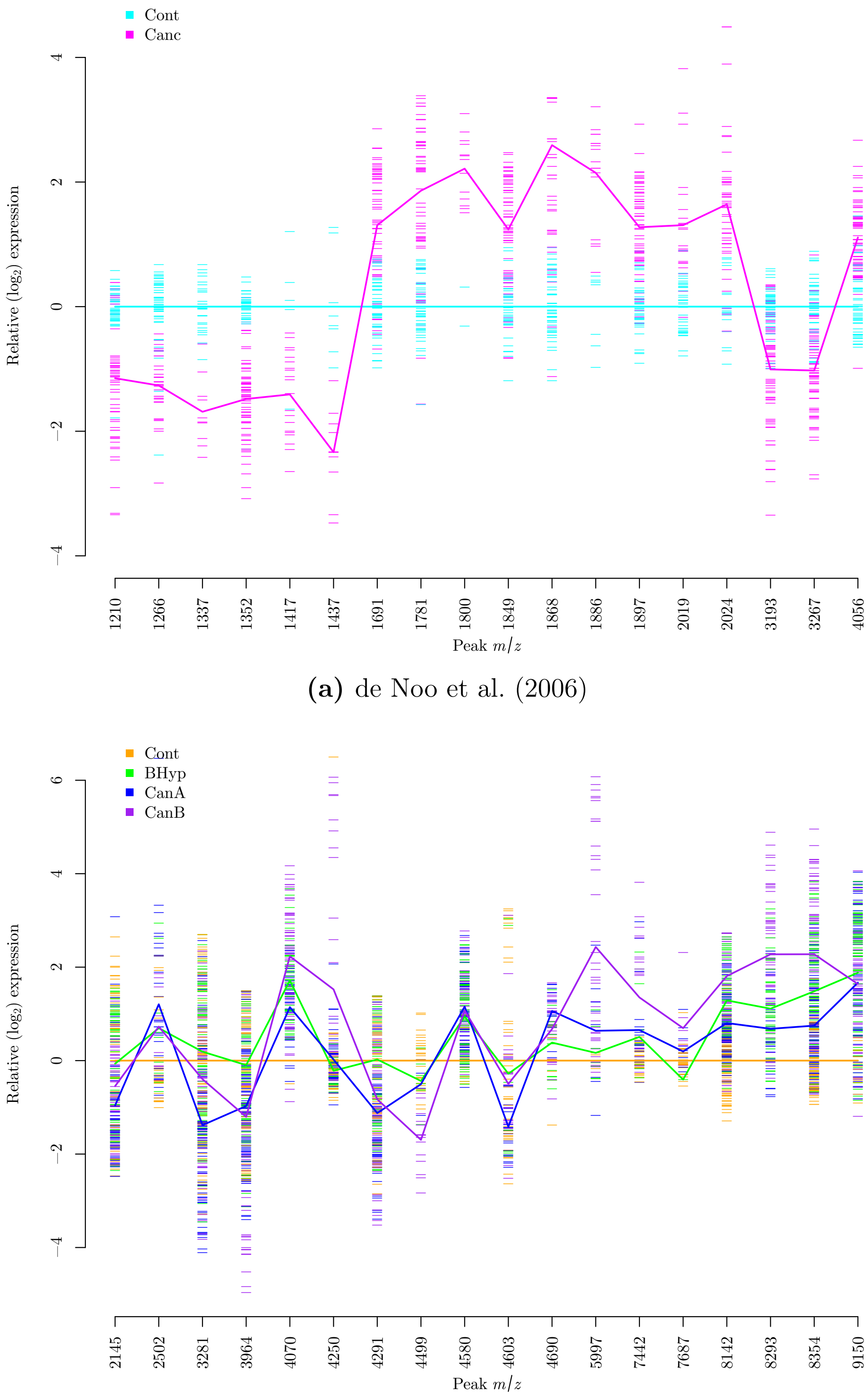


Figure 4.13: Volcano plot for the Adam et al. (2002) dataset for group differences peak expressions; adjusted p-value vs. fold change on the $\log_2$-scale. Missingness observed for each peak is indicated by rectangle fill adjacent to point.

with caution, according to the known problems with the SELDI TOF-MS platform (McLerran et al., 2008a,b).

Figure 4.14 provides a visualisation of relative peak expression on the $\log_2$-scale for disease groups relative to the control group in the de Noo et al. (2006) and Adam et al. (2002) datasets based using the OLS models fitted. Only peaks reaching statistical significance and fold-change threshold are pictured. The down-regulation of lower m/z peaks and up-regulation of higher m/z peaks in the de Noo et al. (2006) dataset discussed previously is particularly evident in Figure 4.14(a). Proteins of



(a) de Noo et al. (2006)

(b) Adam et al. (2002)

Figure 4.14: Parallel plots of peak expression on the $\log_2$-scale relative to the model-estimated control group effect for each peak identified in the volcano plot in Figures 4.12 and 4.13.

most interest in Figure 4.14(b) are those showing a difference in peak expression between the control and hyperplasia groups from the cancer groups.

The asthma1 and asthma2 datasets were regressed using LME models with fixed overall mean and group effects, and random intercepts for mothers to account for repeated measures. For the asthma1 and asthma2 datasets, the fold differences were calculated between the two experimental groups (male and female births).

The asthma1 and asthma2 peak expression datasets uncovered no statistically significant peaks controlling for a FDR at 0.05 using Benjamini & Hochberg; however five and eight peaks, respectively, had fold-changes reaching the one and half fold-change threshold. These results are given in Appendix C in Figures C.8 and C.9 and Tables C.5 and C.6. Two peaks surpassing the fold-change threshold in the asthma1 dataset and five in the asthma2 dataset are highly 'under-observed', and with the assumption of missing at random in doubt, the large fold-change is likely to be due to missingness. The fact that no significant peak expressions were found using the asthma2 dataset is reassuring, as this dataset is not expected to find any peaks with differentiating group expression because of the serum handling issues degrading any true biological signal.

4.3 GC mice: the effect of missingness on statistical inference

The large amount of missingness in the GC mice data (and MALDI/SELDI TOF-MS data generally) is a non-trivial problem that requires serious consideration and further research. How missingness affects the previous inference from the analysis is considered here in two ways: firstly, whether missing values are significantly associated with experimental covariates and secondly, how parameter estimates change in the linear models when missing values are imputed. Together, the two perspectives will help inform whether the potential biomarkers identified in the GC mice data with LME models are likely to be reliable.

Missing at random (MAR) is usually considered a requirement for unbiased estimates in the LME model (Saha and Jones, 2005), also referred to as ignorability requirements (Heitjan and Basu, 1996). The definition of MAR requires the non-observation of values to be independent of the variable in which the missingness occurs. However, the missingness may be related to another factor and still be considered MAR: this is why MAR is a weaker condition than missing completely at random (MCAR). When missingness is unrelated to any measurable factors, it is categorised as MCAR (Rubin, 1976).

Peak expression values can be missing for a variety of reasons but for each individual missing value, the true cause is not known. The missing value can be a result of a protein not existing (or existing at high enough concentration) in the sample to have desorbed from the MALDI/SELDI chip, the protein expression does not meet the algorithmic peak detection threshold, or, it is missing completely at random. It is well established that missingness in peak expression is largely influenced by the abundance of the protein, i.e., proteins that do not meet the detection threshold (Karpievitch et al., 2012).

The average peak expression is plotted against the proportion of missing values for all $P = 159$ peaks in Figure 4.15, demonstrating peaks with a lower abundance of peptides are associated with more missing values. Such a relationship implies that neither MCAR or MAR are suitable assumptions. As a result, the parameter estimates of the LME models on the GC mice data in §4.2.1 may be biased.

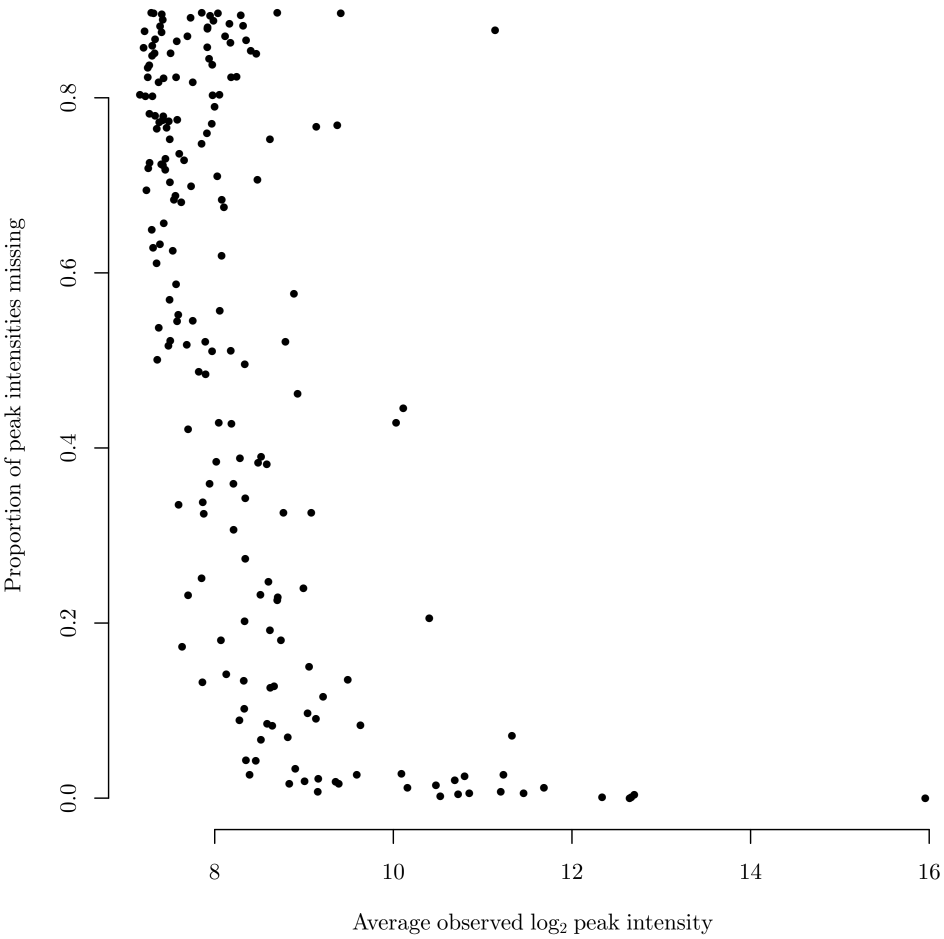


Figure 4.15: Scatter plot of the mean $\log_2$ expression against proportionate missingness for each peak in the GC mice dataset.

In examining the effect of incorrectly assuming MAR in the LME models on the GC mice data in §4.2.1, some logical implications might be considered. If it is assumed all missing values are due to expression below detection limits, the parameter estimates are biased upwards. However, if the missing values are unrelated to disease group, the differences between disease group parameter estimates will be roughly bias free, as all disease groups will have overestimated mean expressions equally. In the scenario where disease group has an effect on the occurrence of missing values and the majority of missing values are a result of having sub-threshold expression, the differences between disease group parameter estimates will be under reported. With these considerations in mind, the following sections attempt to determine the extent of the bias, and the subsequent influence on the statistical inference made in §4.2.1.

4.3.1 Missing peak expression as an outcome

Investigated here is the missingness observed in the GC mice peak expression data. The association of experimental group and chip with missingness is explored here. In fitting a statistical model to each peak expression vector dichotomised as missing or non-missing and controlling for the effect of (mean) peak expression, the relationship between experimental effects and the probability of missingness is examined. To perform such an analysis, generalized linear models (GLMs), utilising general estimating equations (GEEs) to account for correlated observations, are applied.

The GLM and GEEs

A GLM assumes the expected value of the response, μ_i, for observations $i = 1, \ldots, n$ can be modelled by a linear combination of predictor variables and model coefficients, $\mathbf{x}_i^T\boldsymbol{\beta}$, where the expected value may be modified by a link function, g (Nelder and Wedderburn, 1972). The expected value of the outcome is therefore deduced using the inverse of the link function,

$$g\left(\mu_i\right) = \mathbf{x}_i^T\boldsymbol{\beta} \Leftrightarrow \mu_i = g^{-1}\left(\mathbf{x}_i^T\boldsymbol{\beta}\right).$$

Here, the binomial distribution is employed to model the binary outcome of 'missingness'. Peak expressions can be considered missing or not and the expected mean missingness on the set of predictive variables for observation i is μ_i. The natural choice of link function is the logit link, so that $g\left(\mu_i\right) = \ln\left(\frac{\mu_i}{1-\mu_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta}$.

By fitting this GLM, it can be inferred whether or not the missingness in peak expression is related to experimental factors. If experimental factors are related to the missingness, it may be concluded the assumption of MAR is unreasonable. Should this be the case, future imputation models need to address this complexity. The GLM assumes independent observations, which is not the case for the GC mice dataset. The use of GEEs (Liang and Zeger, 1986; Zeger and Liang, 1986) allows adjustment to the standard errors in the model to account for correlation between observations. The expected value of the observations takes a vector representation, $\boldsymbol{\mu}_i$, to handle this clustering of observations, y_{ij}, for subjects $i = 1, \ldots, M$, each with repeated observations $j = 1, \ldots, n_i$.

Generally, a working correlation matrix, $R(\boldsymbol{\alpha})$, needs to be pre-specified to account for the correlation between observations within a subject. This working correlation can be specified in a functional form with parameters $\boldsymbol{\alpha}$ to be estimated by the model. In special cases, the correlation may be known and thus fixed parameters used. A working correlation could take the form as previously illustrated by the heatmaps in Figure 4.8. However, an exchangeable correlation structure within mouse was implemented for simplicity since a misspecified working correlation matrix still allows asymptotically unbiased estimation of $\boldsymbol{\beta}$ (Wang and Lin, 2005).

GEEs and the respective solutions for $\boldsymbol{\beta}$ are estimated using (Wang and Lin, 2005),

$$\sum_{i=1}^{M} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \Lambda_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0},$$

where Λ_i is a function (including a scale parameter) of the working correlation matrix, $\boldsymbol{y}_i$ and $\boldsymbol{\mu}_i$ are the observations and marginal means associated with the repeated measurement on subject $i = 1, \ldots, M$, respectively. The partial differentiation matrix $\frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}}$ is informed by the choice of the link function, specifically the inverse of the specified link function as a function of $\boldsymbol{\beta}$.

GEEs are not a likelihood-based method, but hypothesis testing using Wald statistics under the assumption of roughly Gaussian means is used. Wald tests of group and chip significance on the binary outcome of missingness are made using GEE GLMs controlling for the mean peak expression unless otherwise stated (because of data limitations). The results focus on the potential biomarkers highlighted in §4.2.1 and the corresponding Wald tests are summarised in Table 4.2.

There were five peaks with insufficient missingness to fit the proposed model. It is unlikely this small level of missingness would have materially biased the estimation of the true mean expression parameters made in the LME models (§4.2.1). The peaks in question are at 6821, 8533, 8831, 16030 and $17458m/z$, all with missingness below 3.5%.

A clear pattern observed from Table 4.2 is that mouse group has a highly significant relationship with missingness where a GLM model was able to be fitted. However,

Table 4.2: GEE GLM modelling of the binary outcome of missingness in the GC mice peak expression dataset using the predictors of disease group and chip.

	Peak expression		p-value[†]	
m/z	Available	Missing	Group	Chip
6602	1008	72	<0.001	0.810
6821	1072	8	N/A[††]	N/A[††]
7412	885	195	<0.001	0.150
7806	692	388	<0.001	0.704
8337	599	481	<0.001	0.688[♮]
8533	1062	18	N/A[††]	N/A[††]
8607	862	218	0.004[♮♮]	0.333[♮♮]
8831	1053	27	N/A[††]	N/A[††]
8867	127	953	0.004	<0.001
9305	858	222	<0.001	0.279
12161	873	207	<0.001[¶]	0.001[¶]
13648	988	92	<0.001	0.134
14421	618	462	<0.001	<0.001
14836	729	351	<0.001	<0.001
16030	1044	36	N/A[††]	N/A[††]
17458	1059	21	N/A[††]	N/A[††]

[†] All p-values are adjusted using the Benjamini & Hochberg method for a FDR of 0.05. [††] Missingness not prevalent enough to fit a binary outcome GLM. [♮] No missing values were observed for groups FFIL6 and IL6. The p-value relates to a hypothesis test of at least one of the remaining mouse group means of missingness differs from the other groups, ignoring that the true mean missingness for FFIL6 and IL6 mice are not likely to be consistent with the hypothesis of equal mean missingness for all mice groups based on no missing values being observed for those groups. Table C.2 contains the missingness across groups and chip for reference. [♮♮] No missing values were observed for group IL6. The p-value is calculated as outlined in [♮]. Table C.2 contains the missingness across groups and chip for reference. [¶] Missing values were not observed for all group and chip strata to allow model estimation of parameters, the model was fitted to the chip and group combinations with missing values.

chip is less likely to share a significant relationship with missing expression values within a peak. It is an important observation that the experimental variable of chip number is not significantly associated with missingness, which gives confidence about the reproducibility of the system across the MALDI chips. As missing values are likely to, on average, be associated with disease group, missingness could be informative of disease status if it were not known to experimenters. Whether missingness is informative in the prediction of unknown disease status, at a mouse level, will be investigated in Chapter 6.

Where either chip or disease group are related to missingness in Table 4.2 using the GLM, the parameter estimates for mean peak expression in §4.2.1 are likely to be highly biased and should be interpreted with care. Using the information from Table 4.2 with Figures 4.10 and 4.11, the peaks at 6602, 6821/13648, 8607 and 17458m/z are the best candidates for biomarkers as they are the peaks with the largest fold-changes and either,

i) had a very small number of missing values to be confident of fold-change estimates, or,

ii) had missing values that were *not* associated with chip but were importantly associated with group, suggesting the true fold-change might be underestimated.

This work is consistent with that of the results of Pun (2014) using the GC mice data used here. Using a random selection of peaks, a Bayesian approach was used to model the missing values. A statistically significant relationship between missingness and peak intensity was found for some peaks. However, the modelling highlighted that an assumption that all missing values are below the detection threshold is too strong, as not all modelled missing values were below the minimum expression seen for each peak. For the peaks where it was shown the missing values are likely a result of undetected signal, the assumption of MAR is invalid, which results in biased parameter estimates when naively modelled. However, an interesting finding of Pun's work is in those situations, the parameter estimates of mean peak expression are severely biased but the factor of interest (group) is largely unaffected. This result held irrespective of the proportion of missing values within the peak.

4.3.2 Linear models with imputed data

To investigate the effect of imputation on model parameter estimates and to help assess the validity of the inference about the potential biomarkers assuming MAR expression (§4.2.1), Figure 4.16 provides a comparison of the potential biomarker parameter estimates from the LME models where missing values are ignored (assuming MAR) and where the missing values are imputed using kNN. Estimates from the LME models, as specified in Equation (4.6), of the group effects for FF, FFIL6, FFStat3 and IL6 are shown, relative to the WT group. Open circles represent the parameter estimate using the missing data and closed circles are the estimates for the kNN imputed data. For each peak, the two corresponding estimates are connected by a vertical line. As to be expected, there is little difference between estimates using the two approaches when there is small proportion of missingness. However, almost uniformly, the estimates migrate towards zero when the imputed data are used.

The two potential biomarkers with the highest levels of missingness identified by the original LME models (which ignored missing values) having a fold change above 1.5 and a significant group effect, were the peaks at 8337 and 8867m/z (45 and 88% missing, respectively). Figure 4.16 shows a large migration of the parameter estimates towards zero for these peaks and Figure 4.17 shows they no longer have fold changes above 1.5 as a result. The remaining peaks identified by the LME models ignoring missing values, showed minor changes in parameter estimates because of the low proportion of missing values.

The curious peak at 7806m/z that had up-regulated expression for the WT group showed very little change in parameter estimates when kNN imputed data were used (Figure 4.16). The missing values for this peak were almost exclusively not in the WT group and the imputed values were of moderate expression, relatively, in the remaining groups.

A comparison of the parameter estimates in all $P = 159$ LME models where missing values are ignored (assuming MAR) and $P = 159$ LME models where the missing values are imputed using kNN is provided for reference in Figure C.7 in Appendix C. The peaks are ordered from left to right in increasing proportions of missing values in each peak. The peaks highlighted in Figure 4.16 retain the surrounding black boxes for ease in identification.

A volcano plot to summarise the identified potential biomarkers using LME modelling on the kNN imputed data can be seen in Figure 4.17. The peak at 14836m/z, in addition to the peaks at 8337 and 8867m/z, no longer have a fold changes 1.5 or greater on the log-scale. The dilution of this difference in group expression is due to imputation of 32% of expression values in this peak. As discussed previously, the kNN impute method is unsupervised (does not use group information) and is likely to produce conservative estimates of group differences for peaks with substantial missingness. Other peaks with a large amount of missingness, such as 7806, 9305 and 14421m/z (with missing value proportions of 36, 21 and 32%, respectively), did maintain fold changes above 1.5.

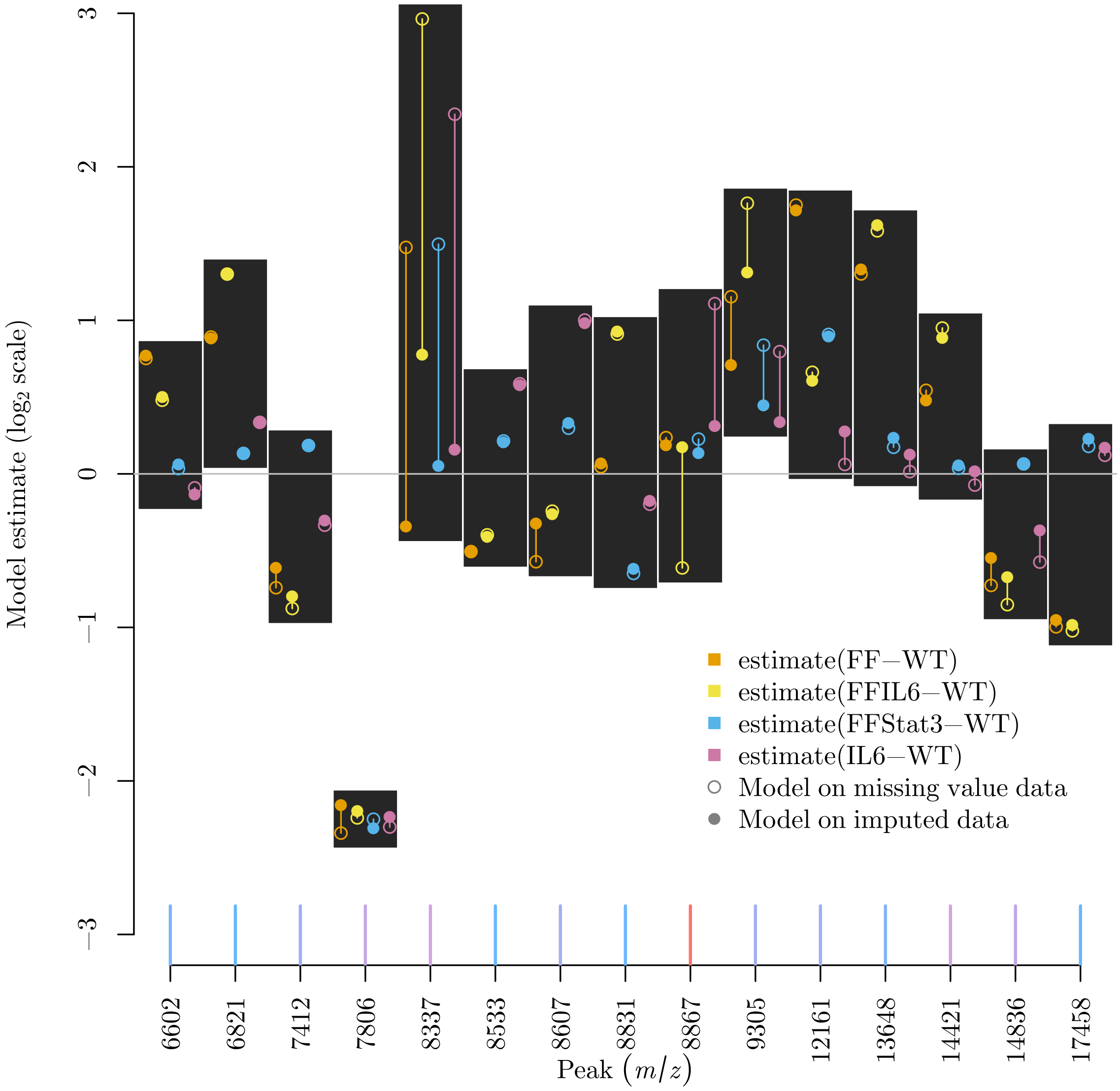


Figure 4.16: Parameter estimates for identified potential biomarkers, relative to the WT group, for the LME models when missing values are ignored and when the missing values are kNN imputed. The ticks on the x-axis coloured on the scale $[blue, red]$ denotes the proportion of missing values in the peak from no missing values to completely missing.

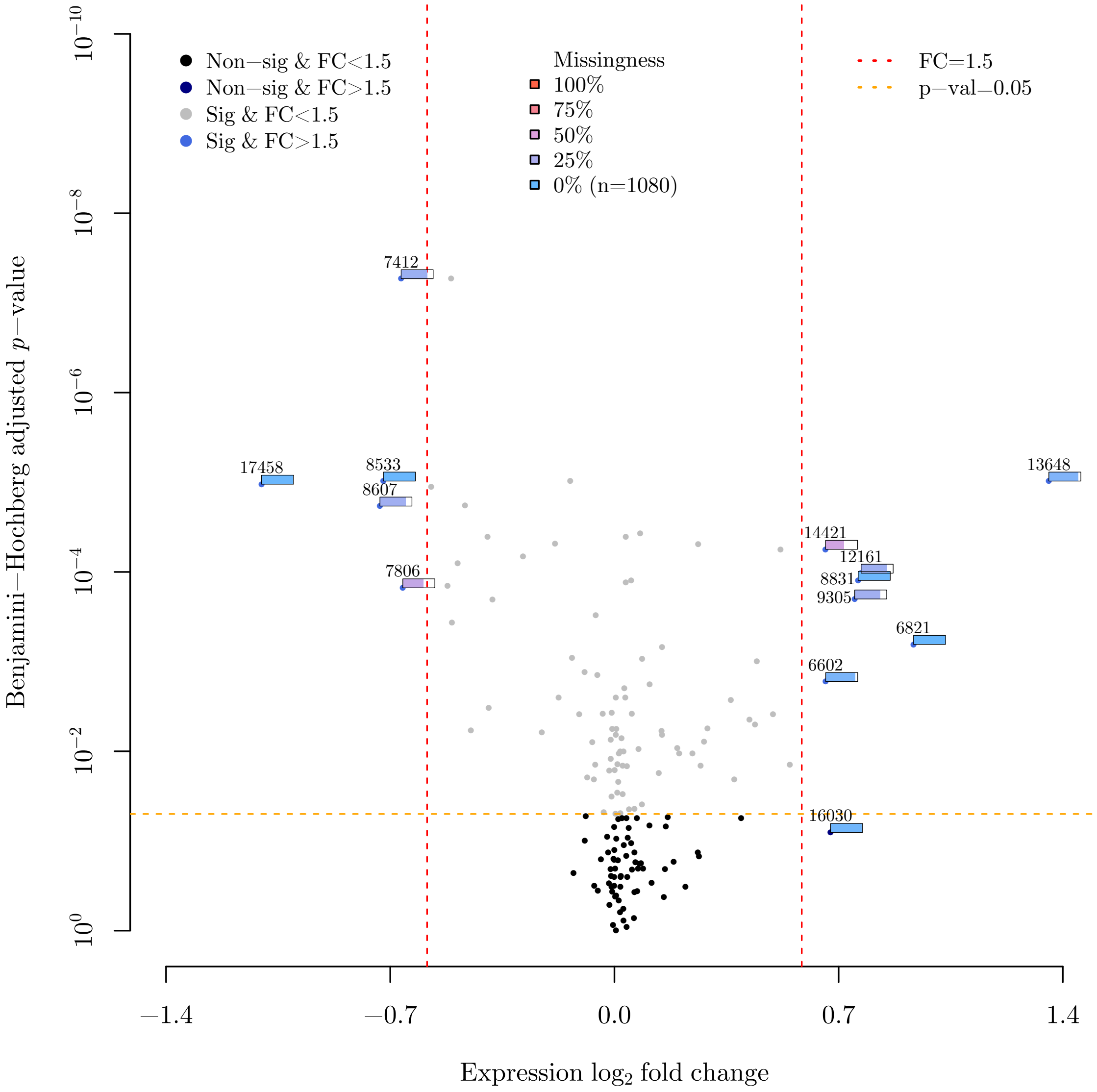


Figure 4.17: Volcano plot of phenotype group differences for each peak of the GC mice expression data with kNN imputed data; adjusted p-value vs. fold change on the $\log_2$-scale. Missingness observed for peaks with fold-change greater than one and a half is indicated by rectangle fill adjacent to point.

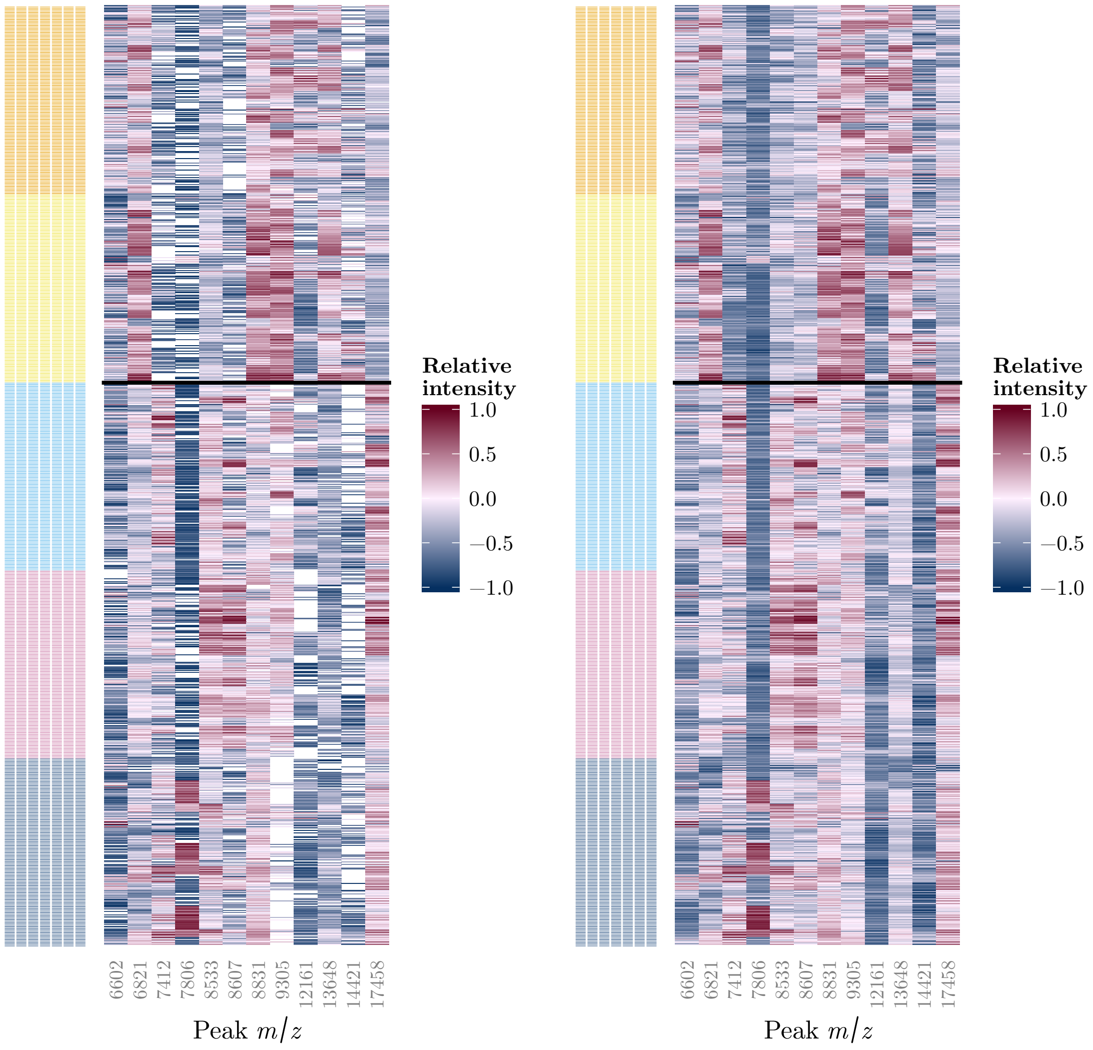
Figure 4.18 shows the expression data for the potential biomarkers identified when using the kNN imputed data. Figure 4.18(a) presents the non-imputed expression data and Figure 4.18(b) presents the imputed data. The horizontal black line divides the peak expression of GC and control mice and within these groupings, the spectra are ordered by chip. For example, the top third of the FF group expression values in the plot correspond to the FF spectra from chip 1. The peaks at 6821, 8533, 8607, 13648 and 17458m/z show the greatest differential expression between the GC and control groups. The differential expression of the WT mice to the other groups in the peak at 7806m/z is highly visible from this figure but it should be noted the higher mean expression of WT mice is highly influenced by a subset of these mice. For this reason, this peak is a poor disease status biomarker candidate as it is likely a latent variable is causing this differential expression in a subset of the WT mice. The peaks at 8533 and 8607m/z show some differential expression within the control mice, with the WT mice having a relatively lower peak expression than the FFStat3 and IL6 groups.

Before a final set of biomarker candidates are selected from the information garnered in this chapter so far, an application of the newly developed methods of surrogate variable analysis and remove unwanted variation are applied to the GC mice data in an attempt to gain further insight on suitable biomarker candidates.

4.4 Unknown and unwanted variation

In proteomic experiments, in addition to known experimental factors, there may be additional unknown experimental factors. Known experimental factors (such as chip number) can be removed using linear models, however new methods to additionally remove unknown experimental effects and variation can be used. One widely used method is Surrogate Variable Analysis (SVA; Leek and Storey, 2007; Karpievitch et al., 2009; Desai and Storey, 2012) that will be explored in the next section. A related method to remove unknown experimental effects, called Remove Unwanted Variation (RUV; Gagnon-Bartsch and Speed, 2012; Jacob et al., 2012), from the microarray literature is also considered in the following section. The aim of these methods is to provide a clearer picture of the variables of interest (i.e. disease classification) by removing artefacts that may not be known or quantified by the experimenter that influence the analysis. These artefacts are referred to as 'batch effects' or 'unwanted variation'.

Consider the p^{th} column ($p = 1, \ldots, P$) of $Y_{(n\times P)}$ as the observed peak intensities for a peptide p. The outcome of peak expression could be considered as a combination of: a peptide mean expression $\boldsymbol{\mu}_p$ vector; treatment or disease state factors contained in a matrix $\mathcal{X}$; incidental and known experimental factors contained in a matrix $\mathcal{Z}$;



(a) No imputation

(b) kNN imputed

Figure 4.18: Peak expression data as a heatmap for the 12 peaks satisfying a fold change of 1.5 and significant group effect when modelled (a) with no imputation and (b) using kNN imputed data. The $\log_2$ peak expressions are scaled to $[-1, 1] = [\text{blue}, \text{red}]$ as relative intensity within peak. The row colours orange, yellow, light blue, purple and blue depict the group membership of spectra derived from mice in the FF, FFIL6, FFStat3, IL6 and WT groups, respectively.

and unknown factors in a matrix $\mathcal{W}$. Thus,

$$\mathbf{Y}_p = \boldsymbol{\mu}_p + \mathcal{X}\boldsymbol{\alpha}_p + \mathcal{Z}\boldsymbol{\beta}_p + \mathcal{W}\boldsymbol{\delta}_p + \boldsymbol{\mathcal{E}}_p, \tag{4.7}$$

for peptides $p = 1, 2, \ldots, P$ where the parameter vectors $\boldsymbol{\mu}_p$, $\boldsymbol{\alpha}_p$, $\boldsymbol{\beta}_p$ and $\boldsymbol{\delta}_p$ are unknown coefficients.[2] Note $\mathcal{X}$, $\mathcal{Z}$ and $\mathcal{W}$ are design matrices assumed to be constant for all peptides, as the experimental design is constant for each set of observed intensities. If the unknown experimental factors, $\mathcal{W}$, were to become known to the experimenter, regression using Equation (4.7) may be performed to extract the unwanted and incidental factors $\mathcal{Z}\boldsymbol{\beta}_p + \mathcal{W}\boldsymbol{\delta}_p$ from the expression data to analyse group differences only.

4.4.1 Surrogate variable analysis

In essence, SVA is a method for estimating additional, systematic variation in the residuals obtained from a linear regression model fit. Using the known experimental variables from Equation (4.7), the estimated residuals $\hat{\mathbf{r}}_p$ for each peak $p = 1, 2, \ldots, P$ are given by,

$$\hat{\mathbf{r}}_p = \mathbf{y}_p - \hat{\boldsymbol{\mu}}_p - \mathcal{X}\hat{\boldsymbol{\alpha}}_p - \mathcal{Z}\hat{\boldsymbol{\beta}}_p, \tag{4.8}$$

where $\hat{\boldsymbol{\mu}}_p$, $\hat{\boldsymbol{\alpha}}_p$ and $\hat{\boldsymbol{\beta}}_p$ are estimates from fitting a fixed effects linear model. An estimate of $\mathcal{W}$ is then extracted from the matrix of estimated residuals, $\hat{R} = \left[\begin{array}{cccc} \hat{\mathbf{r}}_1 & \hat{\mathbf{r}}_2 & \ldots & \hat{\mathbf{r}}_P \end{array}\right]$. The estimated design matrix $\mathcal{W}$ is used in subsequent linear models to remove ‘unknown’ covariates.

The SVA algorithm

The $\mathcal{W}$ matrix is calculated via a singular value decomposition (SVD; Golub and Reinsch, 1970) of the residual matrix,

$$R = UDV^T. \tag{4.9}$$

Note in Equation (4.9), D is a diagonal matrix with descending values $d_1 \geq d_2 \geq \ldots \geq d_{n_u}$, called singular values and $n_u = \min(n, P)$. The matrices U and V are orthonormal, where the columns of U are the eigenvectors of RR^T and the columns of V contain the eigenvectors of R^TR (Golub and Reinsch, 1970). The matrix $\mathcal{W}$ is estimated as a subset of the columns of U called the left singular matrix. The dimension of the matrix U is $n \times n_u$, D is $n_u \times n_u$ and V is $n_u \times P$. Many texts

[2] To avoid confusion in notation with the previous regression models and the design matrices $\mathbf{X}$ and $\mathbf{Z}$, the matrices $\mathcal{X}$ and $\mathcal{Z}$ are used here.

on SVD have the data transposed as a $P \times n$ matrix assuming $P > n$ or $P < n$; presented here is SVD consistent with data in $n \times P$ form with no assumptions about dimensionality.

The SVD is related to PCA which was outlined at the beginning of this chapter (§4.1.1). Using the SVD of R and the properties of Equation (4.9), consider the covariance matrix of R,

$$\begin{aligned}
S_R &= \frac{1}{n}R^T R = \frac{1}{n}(UDV^T)^T UDV^T = \frac{1}{n}VDU^T UDV^T \\
&= \frac{1}{n}VD^2V^T \qquad \text{as } U \text{ orthonormal} \\
\Rightarrow D^2 &= V^T(nS_R)V \qquad \text{as } V \text{ orthonormal} \\
&= V^T(R^T R)V = V^T V\Lambda \qquad \text{by eigenvector definition} \\
\Rightarrow d_j &= \sqrt{\lambda_j} \;\; \forall j = 1, 2, \ldots, n_u \qquad \text{as } V \text{ orthonormal.}
\end{aligned} \tag{4.10}$$

Equation (4.10) establishes the singular values of R are equal to the square root of the PCA eigenvalues of R. This relationship is utilised as an efficient method to calculate the PCA eigenvectors (used in §4.1.1) and values as the SVD method does not require calculation of the covariance matrix of data which can cause numerical imprecision (Jolliffe, 2005).

A subset of the n_u columns of U that explain a significant amount of (non-random) variation in the residuals are of interest. These are the first h columns of U that have corresponding (ordered) singular values that represent a 'greater proportion of variation than expected by chance' (Leek and Storey, 2007). Significance is determined empirically by comparing the observed singular values of $\hat{R}$ against empirical null distributions of singular values from $\hat{R}$ where the column entries are permuted, referred to as permutation p-values of eigenvalues (Buja and Eyuboglu, 1992). Code in `R` to estimate h was taken from `EigenMS/DanteR` (Karpievitch et al., 2009; Taverner, 2012) but is slightly modified to accommodate data with $n \geq P$ such as for the GC mice dataset and to handle data in $n \times P$ form.

Once $\hat{\mathcal{W}}$ is established via selection of h columns of U, the model (4.7) can be fitted to allow the peak expression data to be analysed with experimental factors, known and unknown, removed. `R` code to achieve expression data with surrogate variables removed is provided in Appendix A.4.

The effect of missingness on unknown covariate estimation

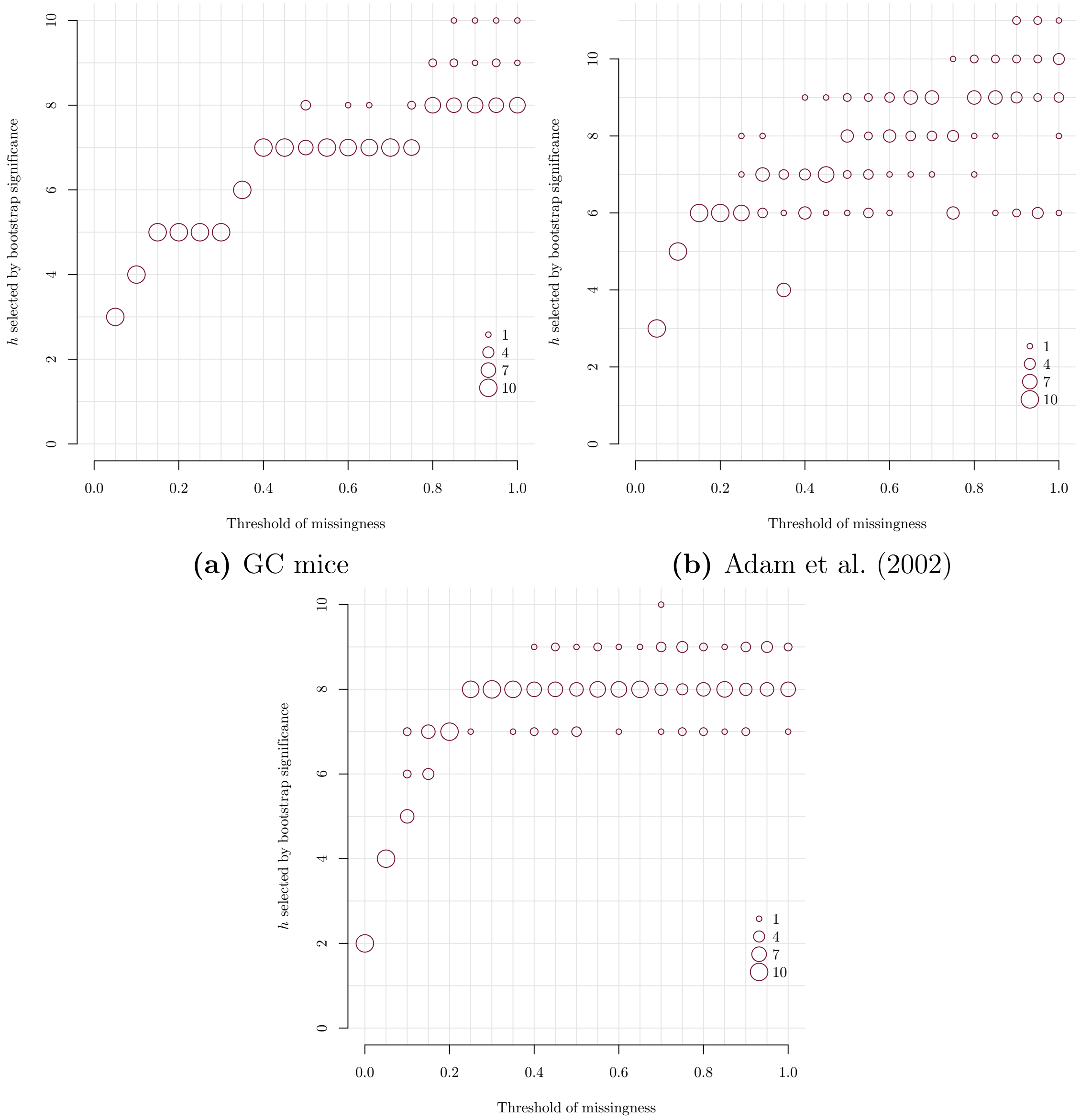
The missing values in proteomic mass spectra data pose a problem for the SVD methods used by SVA and RUV. A simple way to deal with this, as suggested by Karpievitch et al. (2009), is to only include the subset of peptides that exist for all spectra (named *complete* peaks) and subsequently adjust all peak expressions with the SVA estimated unknown experimental factors derived only from complete peaks. This is suggested in the context of liquid chromatography-MS (LC-MS) where expression missingness occurs with smaller probability. As previously presented in Figure 4.15, there is a range of peak missingness proportions in the data. From this plot, it can be observed that very few peaks have complete peak expressions available to perform SVA, as suggested by Karpievitch et al. (2009).

By performing SVA using only the complete peaks on the Adam et al. (2002) and de Noo et al. (2006) datasets, no further insight into potential biomarkers were gained. The statistical power and remaining dimensionality of complete peak data prohibits a sufficiently large h (see Figure 4.19). It is worth noting that SVA will potentially remove variability of estimates (thus affecting p-values) but will not greatly affect mean estimates (and thus fold-changes).

To investigate the effect of increased data availability to the SVA algorithm, imputed values to provide complete peaks were used. The missing values in the residuals after a linear model fit[3] were imputed as opposed to imputing the expression values. This requires fewer assumptions to be made about the missing values. By including peaks below a proportion of missing values threshold, p_{thres}, and imputing random values based on the OLS/GLM's residual standard error for these peaks, the change in the number of surrogate variables estimated with variable p_{thres} can be observed. The use of the random (Gaussian) expressions imputed is justified as it maintains random variation while not adding any systematic experimental effects. By setting p_{thres} to an appropriately small value, issues with non-conformity to MAR will be minimised. No further downstream analysis is undertaken using SVA on using the imputed values, as it is only used here to demonstrate that other, more sophisticated methods need to be developed to be able to successfully perform SVA on MALDI/SELDI TOF-MS peak expression data. Additionally, if further analysis was undertaken using the estimated surrogate variables, an arbitrary cut-off of the threshold p_{thres} would have to be chosen. It is unclear at what point the competing constraints of a low enough p_{thres} to invoke concerns about the MAR assumptions and a p_{thres} high enough to allow enough power to estimate the significant surrogate variables is optimised.

Figure 4.19 shows that including non-complete peaks with imputed data that originally contained a small proportion of missingness, in combination with the complete peak data, a substantial increase in the estimated number of surrogate variables, h, is achieved. For example, for the de Noo et al. (2006) peak expression dataset,

[3] OLS was used for the Adam et al. (2002) and de Noo et al. (2006) datasets. A GEE GLM was used for the GC mice data as it is a 'marginal' model similar to OLS used in SVA methods. A 'conditional' model, like LME, will result in a residual matrix that cannot be interpreted or analysed in the same way, i.e., at the population level. However, using the residuals from the fixed effect component of a LME would be appropriate.



(a) GC mice

(b) Adam et al. (2002)

(c) de Noo et al. (2006)

Figure 4.19: Plot of estimated number of surrogate variables, h, found by the permutation p-values of eigenvalues method for the peak expression datasets. Different thresholds for peak expression missingness determined which peaks were used in the SVA calculation of h. The area of the maroon circle indicates the proportionate number of times h was chosen for that threshold of missingness (each threshold was run for 10 random imputations).

an increase of the missingness threshold to allow peaks that have a proportion of missingness of 0.1 or less changes the estimated h from 3 to 6 when compared to the complete-peak only SVA. The GC mice, Adam et al. (2002) and de Noo et al. (2006) peak expression datasets showed similar trends of a stabilised estimate of h beyond a missingness threshold of 0.5.

4.4.2 Remove unwanted variation

RUV is a similar method to SVA, in that it extracts unexplained variation in the residuals from a linear model fit to estimate the unknown covariates. RUV has been developed for a range experimental situations. RUV 2-step (RUV2; Gagnon-Bartsch and Speed, 2012) and RUV replicates (RUVR; Jacob et al., 2012) are two of these. Not considered here are additional flavours of RUV not applicable to the MS data.

RUV can be considered a more robust method for removing unknown variation than SVA in certain contexts because of its explicit use of expression values that do not vary with the groups of primary interest. RUV also has a significant advantage in being an unsupervised method, as the variable(s) of interest ($\mathcal{X}$) is not used in the estimation of the unknown covariates ($\mathcal{W}$). This is very important if the downstream analysis involves supervised methods for classification. The use of a supervised method prior to classification will result in downwards-biased prediction error, as the outcome variable in classification (disease class) has been used to enhance the predictive data (peak expression data). This is discussed in more detail in Chapters 5 and 6.

RUV using negative controls

RUV2 is very similar to SVA, the major difference being that a subset of the peptides in the residual matrix $\hat{R}$ are used in RUV. Denote this subset of the residuals by $\hat{R}_{\text{nc}}$. The subset of peptides chosen, called negative controls, are peptides that satisfy two criteria:

(1) the peak expression is *unaffected* by the feature of interest (i.e. the factors contained within $\mathcal{X}$) and,

(2) the peak expression is *affected* by the features not of interest (i.e. the factors contained within $\mathcal{Z}$).

This way, the additional variation found in the residuals of the negative control peaks can be estimated without using the factor of interest, $\mathcal{X}$. Therefore the residuals of the negative controls can be calculated by a modified version of Equation (4.8), $\hat{R}_{\text{nc}} = Y_{\text{nc}} - \hat{\boldsymbol{\mu}}_{\text{nc}} - \mathcal{Z}\hat{\boldsymbol{\beta}}_{\text{nc}}$ where $\hat{\boldsymbol{\mu}}_{\text{nc}}$ and $\hat{\boldsymbol{\beta}}_{\text{nc}}$ are matrices composed of $\hat{\boldsymbol{\mu}}_p$ and $\hat{\boldsymbol{\beta}}_p$, respectively, for p in the set of negative control peaks.

Peptides that fit the negative control criteria can either be spiked-in peptides or known peptides in the samples, called house-keeping peptides. If these two types of peptides are not available, the use of empirical negative control peptides can be used.

Complete data are required to compute the SVD and using a complete peak subset of the peak expressions is a limiting factor, similar to SVA. To compound these issues, the number of observed features in MALDI/SELDI TOF-MS are generally an order of magnitude less than that of microarray experiments so the subset of peaks satisfying the (empirical) negative control criteria to undertake SVD make this approach untenable. However, with the richness of replicate spectra available in the GC mice data, an alternative approach suggested by Jacob et al. (2012) is available.

RUV using replicate samples

An alternative version of RUV proposed in Jacob et al. (2012), denoted here by RUV-rep, is another unsupervised method to remove unknown experimental factors. As opposed to using the column-wise approach of RUV2 by limiting the data to negative controls, RUV-rep takes a row-wise approach to utilise variation observed in replicate spectra.

RUV-rep creates a modified peak expression matrix, Y^d, with the same dimensions as Y, where the rows are adjusted using replicate peak expressions. Each row of Y^d is the corresponding peak expression row of Y but with the average of the remaining peak expression replicates for that mouse removed. By making this adjustment, the corresponding design matrix of the factor of interest, X^d, becomes $\mathbf{0}_{n\times P}$. This allows the unwanted variation to be estimated without using the factor of interest.

RUV-rep was proposed assuming complete data which is not the case for the GC mice data. The kNN method of imputation was used to remedy the issue here. In contrast to the situation for SVA, where linear models were fitted on the available data and missing values in the residual matrix were then imputed, RUV-rep does not explicitly employ a linear model to estimate the effects of unknown covariates to generate residuals. As such, initial imputation of the peak expression data is required and therefore a method of imputation that creates the most sensible peak expression data is of primary importance. Figure 4.20 shows the kNN method of imputation yielded sensible complete expression data. Many low expression imputed values were observed which is consistent with the indications that most missing values are likely to be the result of true expression below a detectable threshold.

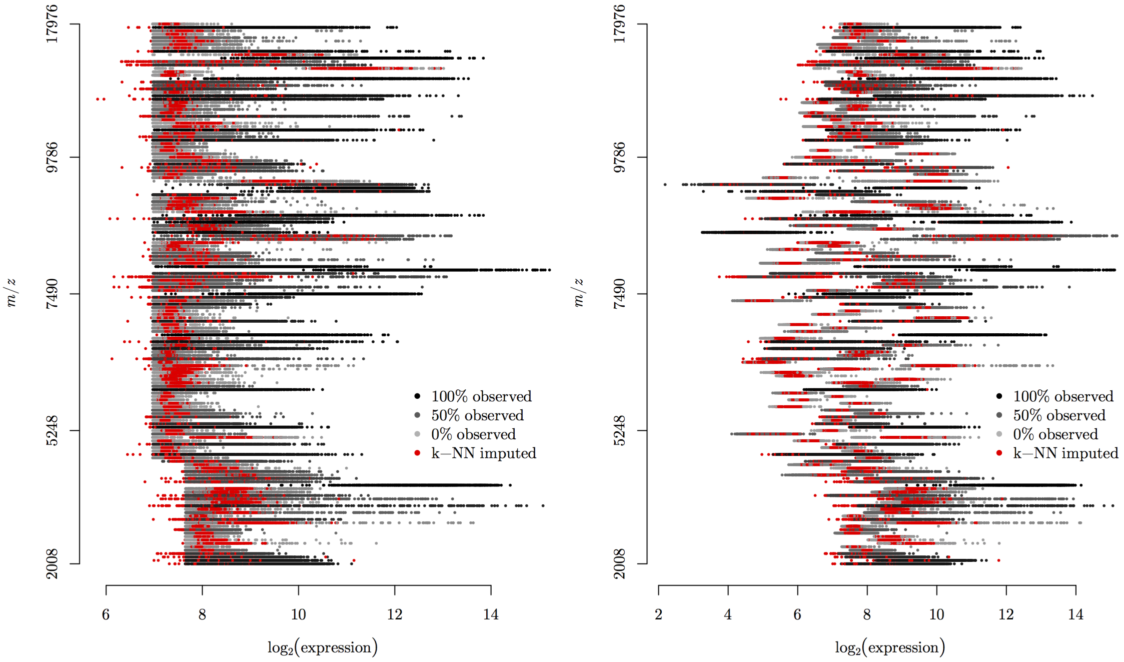
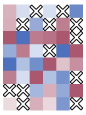
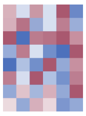
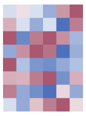
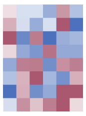


Figure 4.20: Plots of GC mice peak expression data for increasing peak location on the y-axis. Peak expression data with kNN imputed values (left) and the RUV-rep adjusted expression data, Y^{RUV}, using $h = 5$ (right). Points are red if they are imputed values. For observed values, the shade of grey represents the proportion of missingness observed within the peak; the darker the points, the less missingness observed in the peak.

The series of steps to produce the RUV-rep adjusted GC mice expression data, shown in Figure 4.20 on the right, are outlined in Table 4.3. Please note that $\mathbf{\Delta}$ denotes a matrix of all the parameter column vectors $\boldsymbol{\delta}_p$, $p = 1, 2, \ldots, P$ from Equation (4.8). Like SVA, the number of significant unknown covariates, h, is required for RUV-rep. The permutation p-values of eigenvalues method (Buja and Eyuboglu, 1992) to determine the number significant right-singular vectors was employed. The number of significant unknown covariates was estimated as $h = 5$, consistent with the SVA estimate of the number of significant covariates in $\mathcal{W}$ (Figure 4.19(a)).

Figure 4.21 provides an indication of whether the RUV-rep process was successful in removing unwanted variation, while importantly maintaining variation in the factor of interest. Using the principal components that demonstrated the best separation of the GC and control groups (principal components one and three from Figure 4.1), the separation has markedly increased. Unlike Figure 4.1, the means of the group and chip expressions are completely separated between the GC and control groups.

Table 4.3: Pipeline for using the RUV-rep method on the GC mice data.

Start with peak expression data, Y.	$Y =$ 
↓	↓
Impute missing data using `R`-package `impute` to create Y^{imp}.	$Y^{\text{imp}} =$ 
↓	↓
Create $Y^{\text{imp},d}$ with the same dimensions of Y^{imp} but rows are replaced as such: $\mathbf{y}_{ij}^{\text{imp},d} \leftarrow \mathbf{y}_{ij}^{\text{imp}} - \frac{1}{n_i-1} \sum_{j'=1,j'\neq j}^{n_i} \mathbf{y}_{ij'}^{\text{imp}}$ where $i = 1, 2, \ldots, n(= 40)$ mice, $j, j' = 1, 2, \ldots, n_i(= 27)$ is the replicate within i and $\mathbf{y}_{ij}^{\text{imp}}$ is peak the expression vector of the spectrum for mouse i and replicate j.	$Y^{\text{imp},d} =$ 
↓	↓
Estimate $\mathcal{W}\boldsymbol{\Delta}$ using the h right singular vectors of $Y^{\text{imp},d} = UDV^T$. From Jacob et al. (2012), the RUV-rep modified expression data Y^{RUV} that can be used for further analysis is $Y^{\text{imp}} - \hat{\mathcal{W}}\hat{\boldsymbol{\Delta}} = Y^{\text{imp}}\left(I - V_h V_h^T\right)$.	$Y^{\text{imp}} - \hat{\mathcal{W}}\hat{\boldsymbol{\Delta}} =$ $Y^{\text{RUV}} =$ 

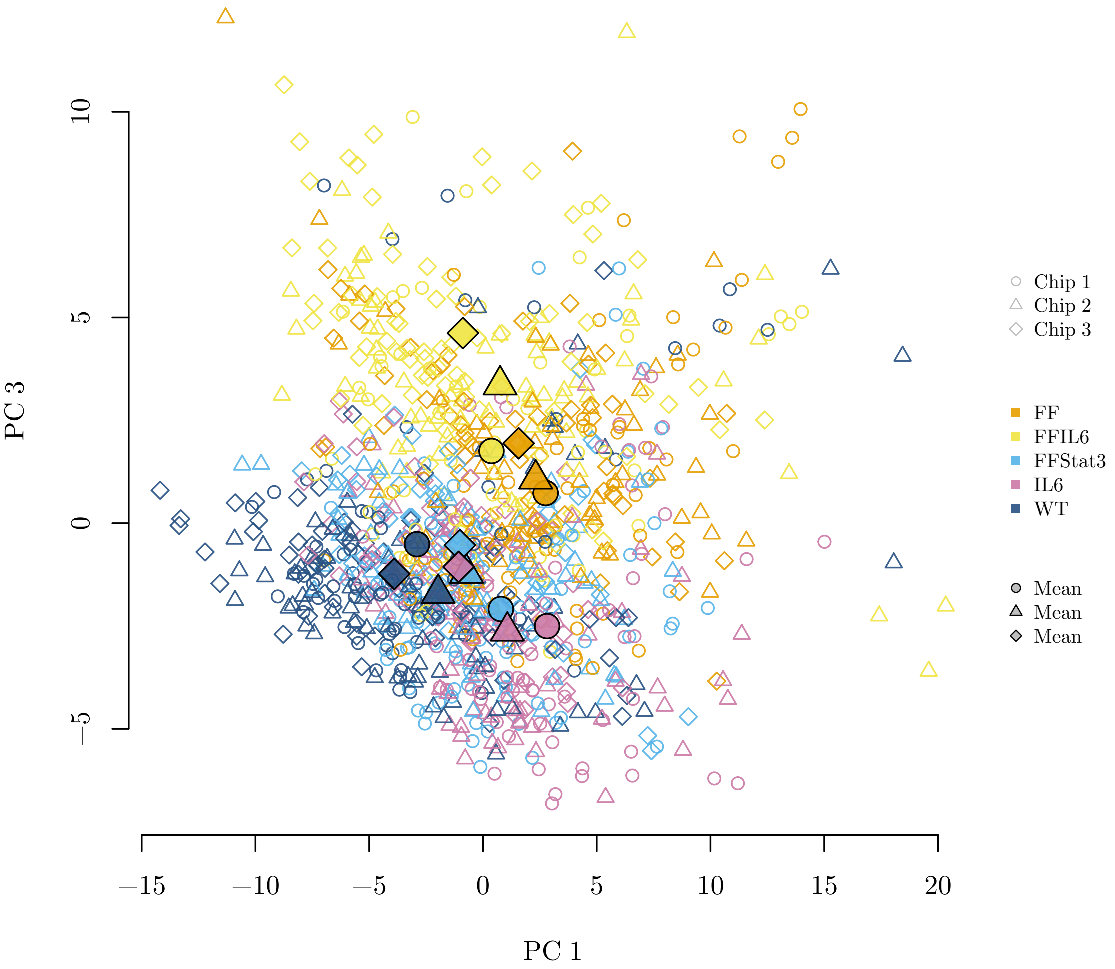


Figure 4.21: PCA plot of RUV-rep adjusted GC mice peak expression data with disease group ($k = 1, 2, \ldots, 5$) and chip ($j = 1, 2, 3$) labelled for principal components one and three. The 1080 PCA points are plotted in a random order irrespective of group membership to avoid a visual bias from plotting points in group order.

Whether or not RUV-rep has materially increased the classification signal between GC and control mice is investigated in Chapter 6. Linear modelling was performed on the RUV-rep expression data, but similarly to SVA, the removal of unknown covariates reduced the variability of peak expression but had little effect on fold-change estimates.

4.5 GC mice biomarker candidate summary

The GC mice dataset has shown the most promise in uncovering proteomic biomarkers for cancer. However, there are significant challenges in the analysis of MALDI TOF-MS data. This chapter has approached the identification of potential biomarkers in a number of ways to build a robust evidence base for conclusions. Table 4.4 summarises the results for the GC mice dataset. Initially, 15 peaks were identified as having a significant group effect and a fold change of 1.5 or greater using the LME models fitted to the data assumed to be MAR.

After the initial selection of biomarker candidates, models to explore possible factors that influence the missing values were fitted to the candidate peaks. Peaks with a small proportion of missing values are not materially affected by peak detection, group or chip effects so it can be concluded the parameter estimates from the LME models are unbiased and the resulting estimates are appropriate for inference. The GEE GLMs, for those peaks with sufficient missingness, demonstrated that not only are missing values influenced by peak abundance and detection but, also by group membership. For the peaks where a group effect on missingness was observed, but not a chip effect, it might be concluded that the LME parameter estimates of differential expression are actually underestimated, as groups with larger proportions of missing values are more likely to have a lower true mean effect than the values estimated. To add to these lines of evidence, peaks that also maintained a FC of 1.5 and significant group effect (after FDR p-value adjustment) for the LME models on the kNN imputed data, remain strong biomarker candidates. Finally, the implementation of the RUV-rep method was used to isolate group signal. It was found that the third principal component specifically provided excellent separation between the GC and control groups. The top 25 contributors (via the absolute values in the third eigenvector) are shown in Table 4.4.

Table 4.4: GEE GLM modelling of the binary outcome of missingness in the GC mice peak expression dataset using the predictors of disease group and chip.

		GEE GLM[††]		Estimated FC	Rank of PC3
	%	Group	Chip	in LME (kNN	contribution in
m/z[†]	observed	effect	effect	imputed data)[♮]	RUV-rep data[♮♮]
6602	93	y	n	1.58↑	19
6821	99	n	n	1.91↑	25
7412	82	y	n	1.59↓	3
7806	64	y	n	1.58↓	
8337	55	y	n	1.11↑	
8533	98	n	n	1.65↓	18
8607	80	y	n	1.66↓	12
8831	98	n	n	1.69↑	
8867	12	y	y	1.02↑	
9305	79	y	n	1.68↑	11
12161	81	y	y	1.71↑	
13648	91	y	n	2.56↑	
14421	57	y	y	1.58↑	1
14836	68	y	y	1.42↓	22
16030	97	n	n	1.60↑	
17458	98	n	n	2.15↓	7

[†] Significant peaks with log fold change ≥ 1.5 using LME modelling assuming MAR (Benjamini & Hochberg adjusted p-values). [††] GEE GLM modelling for missingness as an outcome. 'y' denotes the GEE GLM found a significant relationship between group or chip and missingness, 'n' denotes no statistically significant relationship was found or the number of missing values was insufficient to fit the model. [♮] A blue fold change value denotes the peaks that remained significant (Benjamini & Hochberg adjusted) with log fold-change ≥ 1.5 when LME modelled using the kNN imputed data. A maroon fold change value denotes these criteria were not reached. An up arrow denotes an up-regulated relative GC group expression and a down arrow denotes a down-regulated relative GC group expression. [♮♮] Peak's contribution to the third principal component are ranked from 1 to 159 based on the absolute value of the corresponding entries in the (third) eigenvector. Only the largest 25 contributions are listed.

Using the information presented in this chapter and summarised in Table 4.4, the following biomarkers are recommended for further proteomic investigation[4] in the following order:

- The peaks at 6821, 13648, and 17458m/z. These peaks had a low missing value proportion and the largest fold changes. In addition, all are proportionately high contributors to the third principal component of the RUV-rep data (noting that 6821 is very likely the double-charged version of 13648).
- The peaks at 6602, 7412, 8607 and 9305m/z. These peaks had a significant group effect and a relative fold change of 1.5 or greater for the LME modelling using the kNN data; furthermore the missingness for these peaks was shown to be related to group (and not chip). Additionally these peaks were in the top 25 proportionate contributors to the third principal component of the RUV-rep data.
- The peaks at 7806 and 8831m/z. The 7806m/z peak had a large percentage of missing values (36%) but there was a group effect on the missing values without a chip effect. The 8831m/z peak had a very low count of missing values with a fold change of 1.69 but was not in the top 25 proportionate contributors to the third principal component of the RUV-rep data.
- The peaks at 12161 and 14421m/z. The peak at 12161m/z had a strong fold change of 1.71 but a missing percentage of 19%, and an association with chip. While the 14421m/z peak was the highest proportionate contributor to the third principal component of the RUV-rep data, it had a large of proportion missing values (43%) and there was a significant chip effect on those missing values.

[4] Using the information presented in this chapter and summarised in Table 4.4, the following biomarkers are recommended for further proteomic investigation. A cursory UniProt database (The UniProt Consortium, 2015) search limited to ‘Mus musculus’ (taxon identifier: 10090) provided close mass matches. However, these are unreliable matches as protein identification involves additional measured or estimated protein properties. For example, the protein amino acid sequence and isoelectric point. Such properties could be obtained by tandem mass spectrometry isolating the peptides of interest from the samples.

Chapter 5

Statistical and computational methods of classification

This chapter outlines the feature selection, classification and error prediction methods to be used on the proteomic MS data in Chapter 6. A k-fold cross-validation approach to minimise downwards bias of the predictive error is presented. Standard and non-standard methods of supervised learning are outlined that form the predictive models in the k-fold cross-validation. Different methods of feature ranking are explored as they are an important part of creating discriminatory rules between groups. A novel method to rank feature importance, Pareto Fronts, is compared to other multi-class feature selection methods.

5.1 Supervised learning in the context of diagnostics

For this chapter, n observations in the training data will be considered:

$$(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_n, \boldsymbol{x}_n),$$

with class membership $y_i = \mathcal{C}_k$, $k \in \{1, 2, \ldots, K\}$, and predictive data $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iP})^T$ for $i = 1, 2, \ldots, n$. Using these data, a prediction rule is created for the purpose of taking a future observation, $\boldsymbol{X}_{n+1}$, with unknown group membership, Y_{n+1}, to estimate the observation's group membership, $\hat{Y}_{n+1} \in \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$. In the context of proteomic mass spectra, the y_i are the disease status or classification group and the $\boldsymbol{x}_i$ are the observed intensities of identified peaks in spectrum i.

This form of analysis is usually referred to as *supervised learning* (Mohri et al., 2012) as data of known classification are used to create a model to predict group membership of future observations with unknown group membership. Figure 5.1 is a simple illustration of the two class ($K = 2$) supervised learning problem, however, the multi-class ($K > 2$) case is also considered in the following sections.

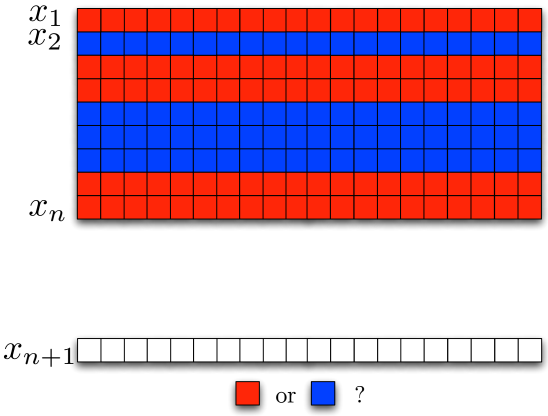


Figure 5.1: The two-class supervised learning problem.

This chapter also includes discussion of issues about feature selection, an important consideration for SnLp-type data, where the number of features is greater than the number of samples or observations in the data, are used in supervised learning.[1] Feature selection aims to find a subset of features that have good predictive value, and at the same time, to remove features that contain no information about the class or group to which the observation belongs.

[1] The standard notation SnLp has been retained here, however P has been used to denote the total number of features and n otherwise denotes the training data sample size from herein.

5.2 Unbiased error prediction

If a predictive model is generated using supervised learning on an entire dataset, then the predictive utility of the model can only be (re-)assessed with observations from the same dataset. The resulting error (sometimes called the *apparent error*; Efron, 1986) will underestimate the true predictive error. The downwards bias of the apparent error is a result of the shared information between the data used to generate the model and the observations used to test the model. To avoid this bias, the observations used to create and test the model should be independent.

To determine the model's ability to make correct predictions, an unbiased error prediction method is of utmost importance. To facilitate independence between observations that estimate the model parameters and the observations that test the model's prediction, the data are randomly split into the training data (model generation, n observations) and the test data (model testing, $N - n$ observations).

The training data are used to create an approximately optimal model with respect to predictive error. The predictive error is the proportion of observations incorrectly classified by the model. Two methods are widely used to make minimally biased (or upwards biased) error prediction within the training data, these are G-fold cross-validation (GFCV) and bootstrapping (Efron and Tibshirani, 1997). Note GFCV is classically referred to as 'k-fold', but to avoid confusion with class membership notation, G has been adopted here. Here, GFCV is used to optimise the model on the training data and re-sampling is used to repeat the (test) error prediction to create distributional results.

The standardised approach to create accurate error prediction of proteomic data in this thesis can be seen in Figure 5.2. The first step is to allocate the available data to the training data and the test data via random allocation. The test data are not called upon until the training data have generated an 'optimal' model.

An approximate ratio of the size of the training to test data is maintained at $\frac{n}{N-n} \approx 2$ for these analyses. There are no established rules regarding optimal training/test data splits, but this is an accepted approach (Kohavi, 1995) and within the generally accepted range of percentage allocation to the training data of 60-80% (Dietterich, 1998; Hastie et al., 2001; Bolton and Bon, 2009). The training/test data split should be a balance between providing enough training samples to generate a stable model and enough test samples to create representative data for error prediction. The error here is the proportion of observations with incorrectly predicted class membership by the model on the test data.

The training data are further randomised into G-folds. The choice of 'G' is somewhat data dependent but 5- to 10-folds is standard and has been shown to be more accu-

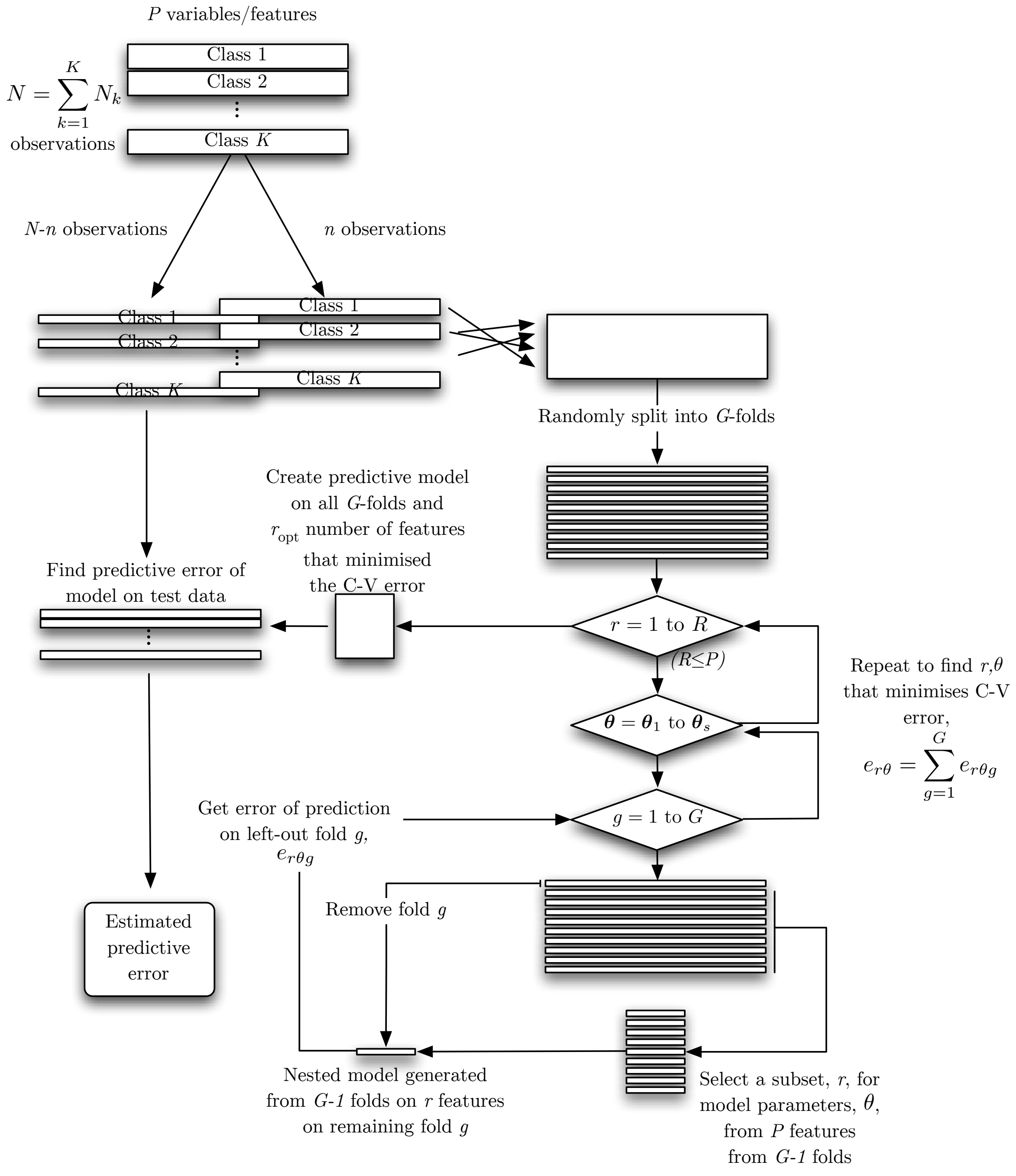

P variables/features
Class 1
Class 2
Class K
$N = \sum_{k=1}^{K} N_k$
observations
N-n observations
n observations
Class 1
Class 2
Class K
Class 1
Class 2
Class K
Randomly split into G-folds
Create predictive model on all G-folds and r_{opt} number of features that minimised the C-V error
Find predictive error of model on test data
$r = 1$ to R
($R \leq P$)
Repeat to find r,θ that minimises C-V error,
$e_{r\theta} = \sum_{g=1}^{G} e_{r\theta g}$
$\boldsymbol{\theta} = \boldsymbol{\theta}_1$ to $\boldsymbol{\theta}_s$
Get error of prediction on left-out fold g, $e_{r\theta g}$
$g = 1$ to G
Remove fold g
Estimated predictive error
Nested model generated from G-1 folds on r features on remaining fold g
Select a subset, r, for model parameters, θ, from P features from G-1 folds

Figure 5.2: Error prediction process using GFCV.

rate than 2- or 3-folds or the more extreme leave-one-out cross-validation, i.e. $G = n$ (Kohavi, 1995). Using $(G-1)$-folds to create a model, then testing the model's error on the remaining fold, allows the suitability of a model to be determined. This is repeated G-times, once for each fold. By summing the errors from each of the G models on the remaining fold, a GFCV error is obtained.

The optimal model using the training data is selected as the model that minimises cross-validated error, $e_{r\boldsymbol{\theta}}$, of the G folds. The cross-validated error is defined as $e_{r\boldsymbol{\theta}} = \sum_{g=1}^{G} e_{r\boldsymbol{\theta}g}$, where $e_{r\boldsymbol{\theta}g}$ is the error of the g^{th}-fold for a set of r features and vector $\boldsymbol{\theta}$ of model parameters. Note that the features are ordered by importance (covered in §5.6) so the choice of $r \in \{1, 2, \ldots, P\}$ is using the top $1, 2, \ldots, r$ features, as determined by feature importance methods. The model parameters $\boldsymbol{\theta}$ are optimised by grid search within the inner loop of Figure 5.2 and are specific to the classification model used; this is covered in §5.3, §5.4 and §5.5.

The G-folds in the cross-validation need to remain constant for each choice of r and $\boldsymbol{\theta}$, otherwise differences in model errors may not be attributable to the variation in parameters but to the variation in the data. Additionally, it is very important to note that while the outer loop in Figure 5.2 is the number of r features used in the model, the r features are ranked and selected only within the innermost loop of Figure 5.2. This design prohibits shared information between the g^{th}-fold, which contributes to the GFCV error, and the other $G-1$ folds. If this were not the case, the predicted error would be downward biased. Another consideration is the independence of replicates. If the observations are derived from the same subjects as replicates, e.g. mice in the GC mice dataset, these common observations should either be exclusively in the training or test data, not both, to maintain independence of the training and test data.

Not all the classification methods require feature selection but will generally benefit from a smaller set of features with predictive value, as opposed to noise, with respect to the group membership of the observations.

Once the optimised model is established on the training data, $(y_1, \boldsymbol{x}_1)$, $(y_2, \boldsymbol{x}_2)$,$\ldots$, $(y_n, \boldsymbol{x}_n)$, the class membership of the test data, $\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+2}, \ldots, \boldsymbol{x}_N$, are estimated by the model. From this, the estimated predictive error is calculated,

$$\hat{e}_{\text{pred}} = \frac{\sum_{i=n+1}^{N} I\left(y_i = \hat{y}_i\right)}{N-n},$$

where $I(.)$ is the indicator function, y_i is the true class membership and $\hat{y}_i$ is the model predicted class of test observations $i = n+1, n+2, \ldots, N$.

To provide information about the variability of the estimated predictive error, the process of allocating the training and test data, allocating the training data to folds, optimising the GFCV model and producing the test prediction error are repeated in the analyses. This process does not add bias to the estimated prediction error as the test data are always independent of the classification model generation. Additionally, the model optimisation using the training data for each re-allocation can be inspected for stability via the number of features selected and similarity of model parameters.

Using the process described in Figure 5.2, the results of the classification methods can be seen in Chapter 6. All analysis was performed in R using self-written functions and R packages outlined in §5.3, §5.4, §5.5. The standardised approach to create predictive error estimates via GFCV model optimisation was additionally written in R.

In the sections that follow, the classification models and feature selection methods are outlined to make this thesis as self-contained as possible.

5.3 Statistical classification

Statistical classification refers to classification models that assume data behave as observations on random variables from probability distributions. As a result, probabilities of class membership can be estimated, which may be of interest in regards to confidence in the predicted class membership.

5.3.1 Linear discriminant analysis

Linear discriminant analysis (LDA) is credited to the paper of Fisher (1936). LDA assumes normally distributed data to discriminate observations between the K possible classes. Consider an observed vector from a P-dimensional Gaussian distribution,

$$(x_{n+1,1}, x_{n+1,2}, \ldots, x_{n+1,P})^T = \boldsymbol{x}_{n+1} \sim \mathcal{N}_P\left(\boldsymbol{\mu}_k, \Sigma\right),$$

where $\boldsymbol{\mu}_k$ is the mean of a class $\mathcal{C}_k$, $k \in \{1, 2, \ldots, K\}$ and Σ is the variance-covariance matrix common to all K classes. In such a context, the class in which $\boldsymbol{x}_{n+1}$ belongs is not known but is of primary interest. This can be framed probabilistically; the probability the data reside in class $\mathcal{C}_k$, given the observed information $\boldsymbol{x}_{n+1}$, is

$$P\left(Y = \mathcal{C}_k | \boldsymbol{X} = \boldsymbol{x}_{n+1}\right) \qquad k = 1, 2, \ldots, K.$$

As the posterior probability $P\left(Y = \mathcal{C}_k | \boldsymbol{X} = \boldsymbol{x}_{n+1}\right)$ cannot be computed directly, Bayes' formula (Ewens and Grant, 2001) can be used, namely

$$P\left(A_j | B\right) = \frac{P\left(B | A_j\right) P\left(A_j\right)}{P\left(B\right)}$$

$$= \frac{P(B|A_j)\,P(A_j)}{\sum_{\forall A_i} P(B|A_i)\,P(A_i)}$$

$$\text{i.e. } P(\mathcal{C}_k|\boldsymbol{x}_{n+1}) = \frac{P(\boldsymbol{x}_{n+1}|\mathcal{C}_k)\,P(\mathcal{C}_k)}{\sum_{i=1}^{K} P(\boldsymbol{x}_{n+1}|\mathcal{C}_i)\,P(\mathcal{C}_i)}.$$

Under the assumption of P-dimensional Gaussian observations, the conditional probabilities $P(\boldsymbol{X} = \boldsymbol{x}_{n+1}|\mathcal{C}_k)$ can easily be computed and the $P(\mathcal{C}_k)$ are the prior probabilities π_k of observing a vector from the group $\mathcal{C}_k$. The predicted class of the observation can now be made with the available probabilities, $P(\mathcal{C}_1|\boldsymbol{x}_{n+1})$, $P(\mathcal{C}_2|\boldsymbol{x}_{n+1})$, $\ldots$, $P(\mathcal{C}_K|\boldsymbol{x}_{n+1})$. Intuitively, classification of $\boldsymbol{x}_{n+1}$ to $\mathcal{C}_k$ is made for

$$P(\mathcal{C}_k|\boldsymbol{x}_{n+1}) > P(\mathcal{C}_i|\boldsymbol{x}_{n+1}) \quad \forall i \neq k, k \in \{1, 2, \ldots, K\}.$$

Linear Discrimination

Given the distributional properties of $\boldsymbol{x}_{n+1}$ are known, if $\boldsymbol{x}_{n+1}$ is an observation from $\mathcal{C}_k$, the corresponding density function is

$$f(\boldsymbol{x}_{n+1}|\mathcal{C}_k) = f_k(\boldsymbol{x}_{n+1}) = \frac{1}{(2\pi)^{P/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}_{n+1}-\boldsymbol{\mu}_k)^T\Sigma^{-1}(\boldsymbol{x}_{n+1}-\boldsymbol{\mu}_k)}.$$

Note that in this scenario it is assumed the variance-covariance matrix Σ is the same for every group $\mathcal{C}_k$.

By considering the creation of boundaries between each class in a pairwise fashion a linear discriminant rule is found (Hastie et al., 2001). The boundaries are created by taking the ratio of the probabilities,

$$\frac{P(\mathcal{C}_k|\boldsymbol{x}_{n+1})}{P(\mathcal{C}_j|\boldsymbol{x}_{n+1})} \text{ for } k \neq j \in \{1, 2, \ldots, K\}. \tag{5.1}$$

If the ratio is greater than one, then the observation $\boldsymbol{x}_{n+1}$ resides in $\mathcal{C}_k$ with higher probability and similarly, if the ratio is less than one, the observation $\boldsymbol{x}_{n+1}$ resides in $\mathcal{C}_j$ with higher probability.

By taking the log of the ratio in Equation (5.1) and simplifying, the classification can be made based on the maximum of the following function,

$$\delta_k(\boldsymbol{x}_{n+1}) = \ln \pi_k + \boldsymbol{x}_{n+1}^T\Sigma^{-1}\boldsymbol{\mu}_k - \tfrac{1}{2}\boldsymbol{\mu}_k^T\Sigma^{-1}\boldsymbol{\mu}_k \quad \forall k \in \{1, 2, \ldots, K\}.$$

The above equation is called the linear discriminant function, as it is linear in $\boldsymbol{x}_{n+1}$ (Hastie et al., 2001). The discriminant function has the geometric interpretation of a linear boundary in the P-dimensional feature space.

Parameter estimation

The parameters π_k, $\boldsymbol{\mu}_k$ (for $k = 1, 2, \ldots, K$) and Σ are usually unknown and need to be estimated. In the proteomics context, each of these parameters is estimated using the training data: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$.

The prior probabilities can be estimated simply as $\hat{\pi}_k = {}^{n_k}/_{n}$, where there are n training observations and n_k training observations in group $\mathcal{C}_k$. As per the method of moments (or maximum likelihood), the estimates of the remaining parameters can be given as $\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{\boldsymbol{x}_i \in \mathcal{C}_k} \boldsymbol{x}_i$ and $\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in \mathcal{C}_k} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)^T$. The `MASS::lda` function in `R` by default uses the method of moments to estimate the parameters.

Unfortunately LDA succumbs to the SnLp problem; Σ will not be invertible, computationally or mathematically, with insufficient observations. Given the number of training observations in each group is $n_1, n_2, \ldots, n_K$ and there are P features, the inequality $n - K \geq P$ must hold for it to be possible to estimate the inverse of Σ.

Other methods of estimating Σ with fewer observations are possible. One such way is to assume features are not correlated. The presence of peaks that derive from differently charged versions of the same protein and the interaction networks of proteins make this assumption questionable.

5.3.2 Quadratic discriminant analysis

Quadratic linear discrimination (QDA) is a result of relaxing the assumptions about the data used in LDA. Now it is assumed the data are,

$$\boldsymbol{X} \sim \mathcal{N}_P(\boldsymbol{\mu}_k, \Sigma_k),$$

where $\boldsymbol{X}$ is in group $k \in \{1, 2, \ldots, K\}$. Note, a different variance-covariance structure is assumed for each class k.

By considering the boundary between two classes as with LDA, a new discriminant function can be obtained with the classification rule of the argument maximum of

$$\delta_k(\boldsymbol{x}_{n+1}) = \ln \pi_k |\Sigma_k|^{1/2} - \tfrac{1}{2} \boldsymbol{x}_{n+1}^T \Sigma_k^{-1} \boldsymbol{x}_{n+1} + \boldsymbol{x}_{n+1}^T \Sigma_k^{-1} \boldsymbol{\mu}_k - \tfrac{1}{2} \boldsymbol{\mu}_k^T \Sigma_k^{-1} \boldsymbol{\mu}_k, \quad k \in \{1, 2, \ldots, K\}.$$

The discriminant function is now quadratic in the data, $\boldsymbol{x}_{n+1}$, and forms a more flexible classification boundary. If the different classes are expected to have different variance structures, a large number of observations per class are required to obtain sensible estimates of the Σ_k.

Parameter estimation

The parameters π_k and $\boldsymbol{\mu}_k$, $\forall\ k = 1, 2, \ldots, K$, are estimated as outlined for LDA. However, the Σ_k for $k = 1, 2, \ldots, K$, are estimated using the training data corresponding to the respective classes. Using the method of moments (or maximum likelihood) the estimates for the Σ_k are $\hat{\Sigma}_k = \frac{1}{n_k} \sum_{\boldsymbol{x}_i \in \mathcal{C}_k} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)^T$.

QDA is even more susceptible to the SnLp problem than LDA. A variance-covariance matrix needs to be estimated for each of the K classes and all are required to be invertible. Given the number of $n_1, n_2, \ldots, n_K$ observations and P features in the data, the inequality $min(n_k) - 1 \geq P$ must hold to for it to be possible to estimate K invertible Σ_k. The use of QDA providing a more flexible boundary has to be balanced against the number of observations and whether enough information is present in the data to reliably fit this more complex structure.

5.4 Extending statistical classification for SnLp problems

In SnLp situations, LDA and QDA require modification to be viable methods of classification because of the limited data. The resultant models are a compromise between the original probabilistic statistical models and algorithmic computational models. These hybrid methods often incorporate integrated or implicit forms of feature selection (§5.6).

5.4.1 Pairwise fusion discriminant analysis

Pairwise fusion discriminant analysis (PFDA) is an example of a classification model based on classical statistical theory with a computational alteration to handle the modern SnLp paradigm. As an analogy, PFDA is to LDA as lasso or ridge regression (Tibshirani, 1996) is to multiple linear regression. PFDA was proposed in Guo (2010) and elements are also outlined in Guo et al. (2010). At the time of writing, no software packages are known to implement PFDA; self-written code to generate the parameter estimates in the PFDA model and classify new observations is provided in Appendix A.5.

Lasso regularised LDA

In the context of SnLp problems it can be beneficial to introduce a Lasso or Bayes (ℓ_1 or ℓ_B term, respectively) to the formulation similar to that described in Tibshirani (1996) and Hastie et al. (2001). The inclusion of these terms penalise estimates of the model parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K, \Sigma)$ that are not considered useful. More specifically, the mean estimates of each class will be penalised towards the mean value of all the classes. As the data are trivially centred and scaled in such a way so that each feature's mean value over all classes is 0, the effect of penalising estimates is to shrink the class means towards 0 in the Lasso penalty function. These adjustments remove certain parameters from the model.

Estimation of group classification and optimisation

PFDA assumes uncorrelated features (zero off-diagonal elements of Σ) and requires penalised parameter estimates, $\boldsymbol{\theta}^* = (\boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*, \ldots, \boldsymbol{\mu}_K^*, \Sigma^*)$. The PFDA discriminant function for $\boldsymbol{x}_{n+1} = (x_{n+1,1}, x_{n+1,2}, \ldots, x_{n+1,P})$ with unknown group membership, is expressed in summation notation as,

$$\delta_k^*(\boldsymbol{x}_{n+1}) = \ln \pi_k - \frac{1}{2}\sum_{j=1}^{P}\left(\frac{\mu_{kj}^* - x_{n+1,j}}{\sigma_j^*}\right)^2 + \frac{1}{2}\sum_{j=1}^{P}\frac{x_{n+1,j}^2}{\sigma_j^{*2}},$$

where σ_j^* are common to all K classes and is the j^{th} diagonal element of the penalised variance-covariance matrix Σ^*, μ_{kj}^* is the j^{th} element of the penalised mean vector $\boldsymbol{\mu}_k^*$ specific to class k.

The term $\frac{1}{2}\sum_{j=1}^{P} x_{n+1,j}^2/\sigma_j^{*2}$ does not depend on k and is therefore a constant for each discriminant function $\delta_k^*(\boldsymbol{x}_{n+1})$, $k = 1, 2, \ldots, K$. The estimated classification can thus be simplified to set $\boldsymbol{x}_{n+1} \in \hat{\mathcal{C}}_i$ for

$$i = \arg\max_k \quad \ln \pi_k - \frac{1}{2}\sum_{j=1}^{P}\left(\frac{\mu_{kj}^* - x_{n+1,j}}{\sigma_j^*}\right)^2.$$

To obtain the penalised parameter estimates, $\boldsymbol{\theta}^*$, a penalty parameter, λ, needs to be optimised. Via iteration over a range of candidate values, the optimised value for λ is chosen where cross-validation error is minimised. As such, λ is a parameter that is placed in an inner loop of the cross-validation classification optimisation process outlined in Figure 5.2.

5.4.2 Regularised discriminant analysis

Regularised Discriminant Analysis (RDA) proposed in Guo et al. (2007), is another variant on traditional LDA for the SnLp scenario. This variant uses the shrunken centroids of Tibshirani et al. (2003) to eliminate unnecessary features.

The first modification RDA makes to LDA is to the estimated variance-covariance matrix, $\hat{\Sigma}$. The estimated variance-covariance matrix is replaced by $\tilde{\Sigma}$ calculated as,

$$\tilde{\Sigma} = \alpha\hat{\Sigma} + (1 - \alpha)\, I_{p\times p},$$

with $\alpha \in [0, 1]$. Using $\alpha > 0$ allows the variance-covariance matrix to be invertible when it might otherwise not be while retaining some of the correlation structure between the features. Such a modification introduces bias into the variance-covariance estimate but can reduce the bias of the discriminant function (Guo et al., 2007).

The second modification relates to the shrunken centroids of Tibshirani et al. (2003). In the discriminant function used to classify observations, the values $\tilde{\Sigma}^{-1}\bar{\boldsymbol{x}}_k = \bar{\boldsymbol{x}}_k^*$ are re-assigned by the relation,

$$\bar{\boldsymbol{x}}_k^* \leftarrow \operatorname{sign}\left(\bar{\boldsymbol{x}}_k^*\right)\left(|\bar{\boldsymbol{x}}_k^*| - \Delta\right)_+,$$

where Δ is a chosen threshold to determine feature inclusion and $(a)_+$ is the positive part of a.

From the two modifications above it can be seen two additional parameters are required for this model, α and Δ. A grid search to find a pair of α and Δ that minimise the cross-validated error is suggested. Similarly to PFDA, these are parameters that are optimised in the inner loop shown in Figure 5.2. RDA has been implemented in `R` as the package `rda` (Guo et al., 2012).

5.5 Computational classification

Computational classification is defined here as a method of classification that does not assume the observations belong to probability distributions. Despite working in a non-probabilistic framework, pseudo-probabilities associated with each estimated classification can be generated from such models. Computational classification is generally algorithmic and optimises decision rules or boundaries via calibration of tuning parameters.

5.5.1 RandomForests

RandomForests (RFs) were proposed by Breiman (2001). RFs are a classification and regression modelling approach that have become feasible with the advent of modern computing power.[2] The use of RFs in this thesis will be in the context of classification and feature selection, both of which will be explained in the coming sections. Before the RandomForest algorithm is outlined here, some of the underlying machinery is explained. RFs are an extension of the classification tree, described below.

Classification trees

A classification tree is built from training data. Starting at a root node with the entire training data, successive nodes split the data to best separate classes. The successive splits terminate at a leaf node which assigns estimated group membership. Figure 5.3 demonstrates this process for a two-class problem using a binary splitting rule, starting with the entire training dataset and recursively splitting the data until only one class exists at the terminal nodes.

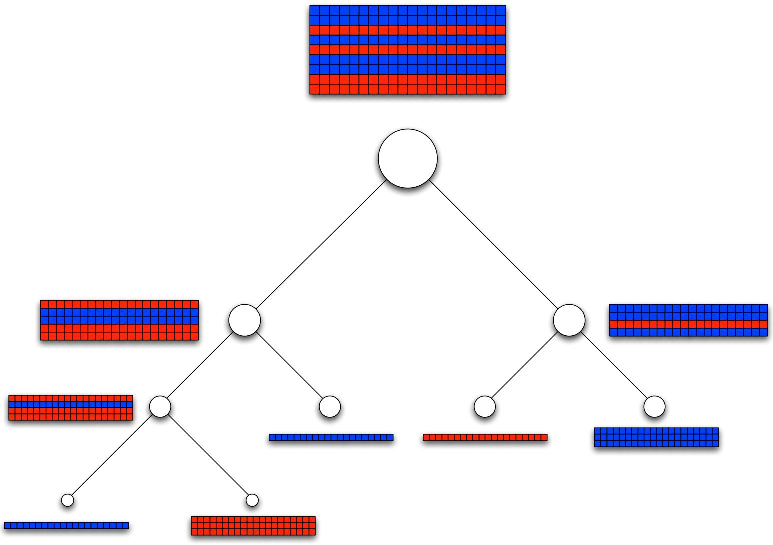


Figure 5.3: An example of a binary classification tree for two-class data, successively splitting data towards terminal nodes. Perfect separation of classes is not always achieved, or necessarily desirable.

For a new observation, $\boldsymbol{x}_{n+1}$, the traversal down the tree using the same splitting rules that created the tree, will classify $\boldsymbol{x}_{n+1} \in \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$.

[2] RandomForests can also be applied to survival analysis and unsupervised classification.

To construct a classification tree as described above, splitting rules to create recursively generated nodes and a stopping rule to determine the final leaf nodes are required.

Classification with RandomForests

RFs are an algorithm that make use of the framework of trees but attempt to minimise their disadvantages.

The RF algorithm works by building N_T trees to make a 'forest' of trees. Unlike traditional classification trees, a random selection of observations and variables are used to make each tree. By only using a subset of features and observations, the algorithm uses information that might not otherwise be included in a single classification tree. This adds stability and is analogous to the use of random perturbation to avoid local optima for the goal of finding global optima.

The construction of each tree in the forest uses randomly sampled observations with replacement, i.e. bootstrapping. The number of random samples m is the same as the number of observations in the data, n. Given sampling is performed with replacement, the resulting sampled data, colloquially termed the 'bag', will likely not include all the original training observations. To determine the expected number of unique observations in the sampled data, consider observation i and let the random variable Z_i be the number of times observation i occurs in the sampled data. Z_i is binomially distributed with the n trials and probability $1/n$ of being selected in each trial. Therefore the probability of observation i being included in the sample is,

$$P(Z_i > 0) = 1 - P(Z_i = 0) = 1 - \binom{n}{0}\left(\frac{1}{n}\right)^0\left(\frac{n-1}{n}\right)^n = 1 - \left(\frac{n-1}{n}\right)^n.$$

Table 5.1 shows for different values of n, the expected proportion of the original observations included in the sampled data. Note that $1 - \lim_{n\to\infty}\left(\frac{n-1}{n}\right)^n = 1 - \frac{1}{e}$. The observations that are left out of generating the tree are utilised in feature importance, these out-of-bag observations will be discussed later.

Table 5.1: Probability of an individual observation being sampled for a tree in a RandomForest.

n	1	2	3	4	10	100	∞
$P(Z_i > 0)$	1	0.75	0.704	0.684	0.651	0.634	$\frac{e-1}{e} \approx 0.632$

The number of features that are considered at each node, q, in constructing a tree in the RF are a randomly selected subset of the feature set. Currently in the R-package `randomForest` (Liaw and Wiener, 2002), the default number of features considered at each node is $q = \sqrt{P}$, where P is the size of the feature set. As outlined in the technical report Breiman (2002), the classification performance of RFs are only sensitive to one input parameter, the selection of q. Increasing q will improve each individual tree's predictive strength, while a decrease in q will decrease the correlation from one tree to another. This trade-off needs to be optimised via cross-validation.

Unlike classification trees, RandomForest trees are built without pruning until terminal nodes are pure. In effect, over-fitting is performed on the bootstrap sample. Figure 5.4 outlines the process of creating a tree in the RandomForest. Like all methods of prediction, over-fitting can occur. Despite this, the variation in feature and data sets in particular nodes seems to offset inherent over-fitting of fully grown trees.

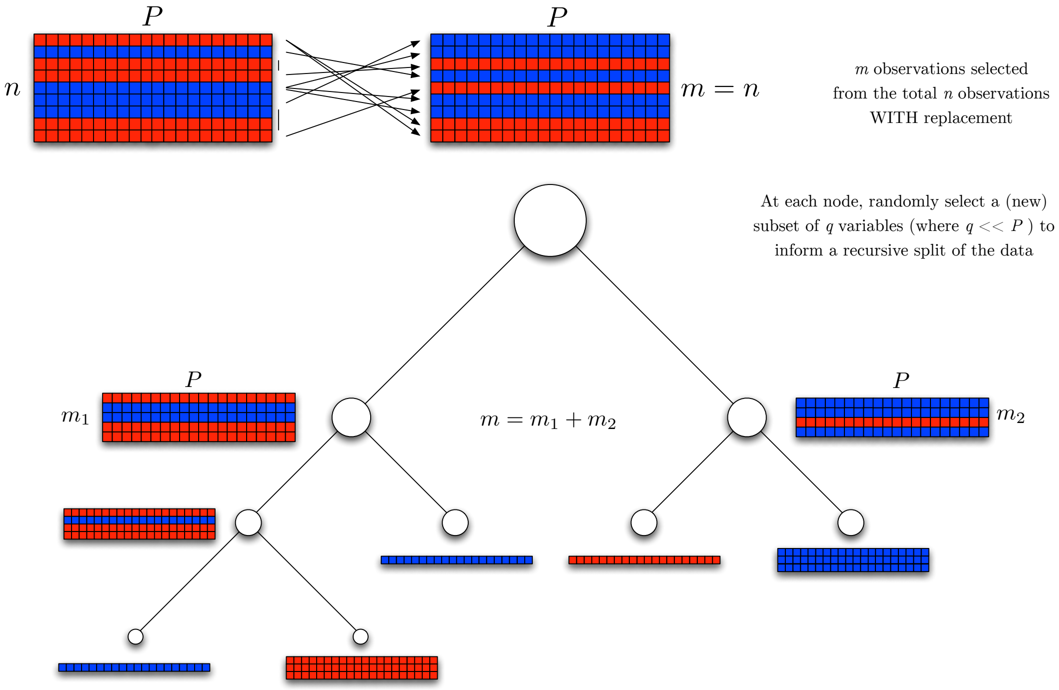

P
n
P
$m = n$
m observations selected
from the total n observations
WITH replacement
At each node, randomly select a (new)
subset of q variables (where $q << P$) to
inform a recursive split of the data
P
m_1
$m = m_1 + m_2$
P
m_2

Figure 5.4: An illustration of the construction of a classification tree for two-class data in a RandomForest.

To generate RF classification, each tree in the forest votes for what an input's classification is, the class with the most votes 'wins'. In all RF analyses in this thesis, N_T has been set to 500 and Gini node impurity (Breiman, 1996) is used.

RFs are a widely used method of prediction and feature importance ranking as they can handle multi-class data easily. RF feature importance takes into account the often complex relationships with other variables, which is valuable especially in circumstances where features are highly correlated. The description of RF feature ranking is found in §5.6.3.

5.5.2 Support Vector Machines

Support Vector Machines (SVMs) are another computational method of classification and when used with a kernel, becomes a non-linear classifier. SVMs have been widely and successfully used since the original paper's description (originally coined support vector networks; Cortes and Vapnik, 1995). SVMs are an extension of optimal separating hyperplanes (Vapnik, 1999) that can be used when the data classes are not perfectly separable. As opposed to using all the data to generate a classifier, SVMs are mainly concerned with the data near the boundary.

SVMs are used to separate two classes but extension to the multi-class case is easily achieved. Initially, an overview of the two-class case will be provided. Once again, training data, $(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_n, \boldsymbol{x}_n)$, are available where $\boldsymbol{x}_i \in \mathbb{R}^P$. However, only two classes are to be differentiated and the classes are designated opposite-signed integer values (Hastie et al., 2001),

$$y_i = \begin{cases} 1 & \text{if } \boldsymbol{x}_i \in \mathcal{C}_1 \\ -1 & \text{if } \boldsymbol{x}_i \in \mathcal{C}_2. \end{cases}$$

Define the hyperplane (or affine),

$$\left\{ \boldsymbol{x} : f(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x} = 0 \right\},$$

and the classification rule,

$$\hat{\mathcal{C}}(\boldsymbol{x}_i) = \operatorname{sgn}\left(f\left(\boldsymbol{x}_i \right) \right). \tag{5.2}$$

Here, $\boldsymbol{x}_i$ is the input for the classifier $\hat{\mathcal{C}}$, which decides the estimated class of $\boldsymbol{x}_i$ based on the resulting scalar value $f(\boldsymbol{x}_i)$. The function f has parameter values $\beta_0, \boldsymbol{\beta}$ estimated by the relationship between the classes (y_i) and inputs $(\boldsymbol{x}_i)$ in the training data.

The optimisation problem of SVMs can be seen in Equation (5.3) below. In essence, Equation (5.3) seeks to find the constant and normal vector, β_0 and $\boldsymbol{\beta}$ respectively, of the hyperplane $f(\boldsymbol{x})$ that separates the two classes, while allowing for some of the training data to exist on the wrong side of the hyperplane for flexibility. The $\xi_1, \xi_2, \ldots, \xi_n$ in the formulation are the distances of the allowed misclassified observations, while γ constrains the total amount of misclassification:

$$\min_{\beta_0, \boldsymbol{\beta}} \quad \tfrac{1}{2} \left\| \boldsymbol{\beta} \right\|^2 + \gamma \sum_{i=1}^{n} \xi_i$$

$$\begin{aligned} s.t. \quad & y_i(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i) \geq 1 - \xi_i && \forall \;\; i = 1, 2, \ldots, n, \\ & \xi_i \geq 0 && \forall \;\; i = 1, 2, \ldots, n. \end{aligned} \tag{5.3}$$

Equation (5.3) is an optimisation problem with a quadratic objective function (to minimise) and linear constraints. The method of Lagrange Multipliers can be employed to create a *dual* program that will also have a quadratic minimising function with linear constraints but the dual program is a simpler (computationally) convex optimisation problem. The solution to the dual program is a solution to the original program (Vapnik, 1999).

Closed form solutions are not available for the dual SVM and a numerical solution requires computation. There are many numerical methods available to solve the program. The `R` package `e1071` (Dimitriadou et al., 2011) is a wrapper package for the `LIBSVM` library of code (Chang and Lin, 2011) and uses decomposition-type methods to compute solutions. Decomposition methods create binary sub-problems using graph theory; the current implementation in `LIBSVM` uses Sequential Minimal Optimisation methods (Fan et al., 2005; Chen et al., 2006). Not only do the sub-problems create tractability for the algorithm, the sub-problems can have analytical solutions providing fast computation (Cherkassky and Mulier, 2007).

Support Vectors and kernels

The 'support vectors' of SVMs are the observations, $\boldsymbol{x}_i$, that correspond to the equations for i where $y_i\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i - (1 - \xi_i)\right) \geq 0$ in Formulation (5.3). This has the geometric interpretation of vectors $\boldsymbol{x}_i$ that lie on the wrong side of the hyperplane decision boundary (where $\xi_i > 0$).[3] These are the observations that shape the hyperplane decision boundary.

The above leads to the concept of non-linear classification SVMs. By enlarging the feature space to make the decision boundary, the hyperplane $f(\boldsymbol{x})$ becomes more flexible. The feature space used so far is $\boldsymbol{x}_i \in \mathbb{R}^P$. This feature space can be enlarged using basis expansions such as polynomials or splines. These basis expansions are functions of the observed data $\boldsymbol{x}_i$ into a larger feature space,

$$\phi(\boldsymbol{x}_i) = \left(\phi_0(\boldsymbol{x}_i), \phi_1(\boldsymbol{x}_i), \ldots, \phi_{M-1}(\boldsymbol{x}_i)\right),$$

where $M > P$. Using this more flexible and enlarged feature space, a new decision boundary is fitted,

$$f(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^T \phi(\boldsymbol{x}) = 0. \tag{5.4}$$

[3] Also corresponds to $\boldsymbol{x}_i$ that lie on or within the margin of the hyperplane.

The underlying concept is instead of fitting a linear boundary in the original space, a linear boundary in the enlarged space is fitted, which when transformed back into the original space is non-linear. The SVM formulation is the same as previously described except the new decision boundary in Equation (5.4) is used with the new classifier, $\hat{\mathcal{C}}(\boldsymbol{x}_i) = \operatorname{sgn}(\beta_0 + \boldsymbol{\beta}^T \phi(\boldsymbol{x}_i))$.

The function $f(\boldsymbol{x})$ can be re-written using a result from the dual SVM program,

$$\begin{aligned} f(\boldsymbol{x}) &= \beta_0 + \boldsymbol{\beta}^T \phi(\boldsymbol{x}) \\ &= \beta_0 + \left(\sum_{i=1}^{n} \lambda_i y_i \phi(\boldsymbol{x}_i) \right)^T \phi(\boldsymbol{x}) \\ &= \beta_0 + \sum_{i=1}^{n} \lambda_i y_i \left\langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}) \right\rangle , \end{aligned}$$

where $\langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}) \rangle$ is the inner product of the transformed data, $\phi(\boldsymbol{x}_i)$ and $\phi(\boldsymbol{x})$. Because $f(\boldsymbol{x})$ only relies on the inner product of $\phi(\boldsymbol{x}_i)$ and $\phi(\boldsymbol{x})$, the transformation $\phi(\boldsymbol{x})$ is not explicitly required, but rather the inner product defined by a kernel function K,

$$K(\boldsymbol{x}, \boldsymbol{x}') = \left\langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \right\rangle .$$

The kernel function is at the discretion of the user. Two of the most used kernel functions are the radial basis kernel and the d^{th} degree polynomial kernel (Hastie et al., 2001; Karatzoglou et al., 2005),

$$\begin{aligned} \text{Radial basis kernel:} \quad & K(\boldsymbol{x}, \boldsymbol{x}') &= e^{-\frac{1}{c}\|\boldsymbol{x}-\boldsymbol{x}'\|^2}, \;\; c > 0, \\ d^{th}\text{degree polynomial kernel:} \quad & K(\boldsymbol{x}, \boldsymbol{x}') &= \left(1 + \langle \boldsymbol{x}, \boldsymbol{x}' \rangle\right)^d, \;\; d > 0. \end{aligned}$$

Although the transformation of the data to the new feature space is not required to be computed explicitly, the polynomial kernel function is generally used to illustrate a expanded feature space. An expanded feature space via the polynomial kernel is demonstrated by finding the finite basis expansions $\phi(\boldsymbol{x}) = (\phi_0(\boldsymbol{x}), \phi_1(\boldsymbol{x}), \ldots, \phi_{M-1}(\boldsymbol{x}))$ for a specified d. The radial basis kernel, however, has an infinite number of basis expansions because of its power series representation.

Choice of kernel

The kernel should ideally be chosen with the data in mind. If the polynomial kernel with $d = 1$ is used, this provides the linear boundary (offset by 1). Using $d = 2$ or 3 provides more flexibility in the margin. Large d can be dangerous for classification problems without knowledge of the data's behaviour and needs to be chosen so that the model does not over-fit the data. This is especially the case for high-dimensional data where the data are likely to be sparse and not require too much deviation from linearity. Potentially, d could be chosen large enough resulting in a decision boundary where every point lies on the correct side of the boundary. But this is unlikely to be useful when it comes to prediction (i.e. when non-training data are used) because the boundary will be too detailed or 'squiggly', compensating for each data point.

The radial basis kernel can be a good choice as it is considered well-behaved because of its 'smoothness' in fit when reduced back to the original feature space. The choice of c dictates how curved the boundary is in the original feature space. A large c dampens the flexibility of the boundary while a small c allows the boundary to behave in a manner that would be comparable to a large d polynomial kernel.

Choice of model parameters and further considerations

The choice of γ and the effect on the margin can be seen by considering the objective function $\frac{1}{2}\left\|\boldsymbol{\beta}\right\|^2 + \gamma\sum_{i=1}^{n}\xi_i$ of Formulation (5.3). As minimisation of the objective function is the primary aim, a very small value of γ will allow non-zero ξ_is without much cost to the objective, which in turn allows a bigger margin. Similarly, a very large value of γ will try to prevent non-zero ξ_is, which in turn creates a smaller margin.

Ringing true to the classical statistical dilemma, the aim is to fit a model which offers a good balance in the bias versus variance trade off. The parameters c or d should be chosen so there is a minimally curved boundary created representative of the two populations in the original space while a sensible choice of γ restricts the misclassification of the training data (support vectors).

Like most computational and non-probabilistic classifiers, optimisation of the model parameters is required. Classically, a grid search of parameters γ and c (or γ and d) is used in the G-fold cross-validation of the training data (without the use of the test data). This adds $n_c \times n_\gamma$ sub-iterations per fold of the cross-validation to estimate the optimal n_c, n_γ combination for the model used.

From testing on the proteomics data, the performance of the radial basis kernel was superior to low-dimensional polynomial kernels in test data classification and will be the kernel function used for all SVMs fitted in this thesis.

SVMs are not invariant to data transformation. This can be observed in the outline of the kernel functions, as features with larger values in magnitude will dominate the returned kernel values. The implementation of `LIBSVM` in the `R` `e1071` package is internally scaled so features dominant in value but not signal do not skew the SVM model.

Multi-class case

To handle more than two classes in the data, $K > 2$, multiple SVMs need to be created.

The simplest method for handling the multi-class case is the one-vs.-all method. This involves creating K models where one group $\mathcal{C}_k$ is compared to all other groups combined, $\mathcal{C}_{k^*} = \mathcal{C}_1 \cup \ldots \cup \mathcal{C}_{k-1} \cup \mathcal{C}_{k+1} \cup \ldots \cup \mathcal{C}_K$. The distances the observation lies from the K classification boundaries, $f_k(\boldsymbol{x}_i)$, are calculated for $k = 1, 2, \ldots, K$. The maximum distance into the classification region of class $\mathcal{C}_k$, opposed to $\mathcal{C}_{k^*}$, corresponds to the predicted class of the observation.

An alternative SVM multi-class classification method is the one-vs.-one method (and is used as the default in `LIBSVM`) that creates $\binom{K}{2} = \frac{K(K-1)}{2}$ models comparing each pairwise combination of classes. The class that *wins* the most pairwise comparisons is the predicted class. Obviously this requires more SVM models for $K > 3$ than the one-vs.-all method but is likely to be a more stable classifier. An improvement on the one-vs.-one method was proposed by Hastie and Tibshirani (1998), called pairwise coupling. This uses the resulting model estimates of pairwise class probabilities, $r_{kk'} = P(\boldsymbol{x} \in \mathcal{C}_k | \boldsymbol{x} \in \mathcal{C}_k \cup \mathcal{C}_{k'})$, to iterate towards a set of probabilities for each class, $p_k = P(\boldsymbol{x} \in \mathcal{C}_k)$, if a solution exists.

5.6 Feature Selection

Feature selection is the selection of variables (peptide peaks) that are important for successful classification of the classes.

A subtle part of feature selection, as opposed to traditional statistics testing parameter differences between groups, is that feature selection should be geared towards differences in individual observations. A feature with a highly significant difference in means (say via a two independent group t-test) may not actually be a good discriminating feature; two means may reach a significance in difference but have large variability around those means and lack predictive power of group membership at the individual level.

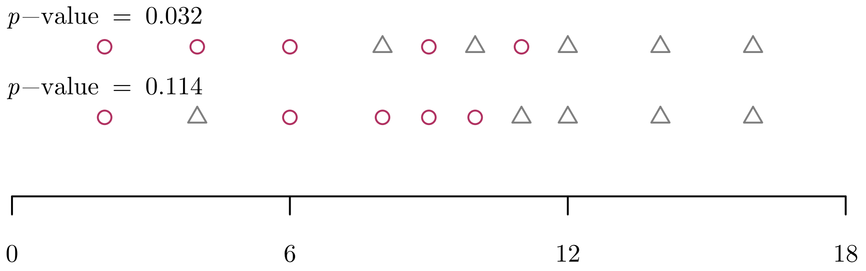


Figure 5.5: The expression for two different features in a two-class problem. The first feature is considered 'more' significant despite having less power to differentiate the two groups. The two groups are differentiated by different plotting characters and colour.

The situation depicted in Figure 5.5 is contrived but illustrates that while there will be a strong positive correlation between features that differ in centrality measure and features with discriminatory ability, these are not an equivalent problem.

This section will outline some of the methods used to determine the importance of features in their ability to differentiate the classes. The feature selection methods proposed are importantly able to handle the multi-class ($K > 2$) case, a requirement for some of the data analysed in this thesis.

5.6.1 Fisher Score

Fisher Scores (FSs) are a simple method to rank feature importance. Even though FSs are an example of a test statistic assessing the difference in class means (§5.6), its inclusion here will be used as a comparison to other feature selection methods. It can be seen FSs are F-statistics when the appropriate constants are applied.

The Fisher Score (Dubitzky et al., 2007) of the j^{th} feature for $k = 1, \ldots, K$ classes, is defined as

$$FS(j) = \frac{\sum_{k=1}^{K} n_k \left(\mu_j - \mu_{kj}\right)^2}{\sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(\mu_{kj} - x_{kji}\right)^2}, \tag{5.5}$$

where n_k is the number of observations in class k, μ_j is the mean value for the j^{th} feature, μ_{kj} is the mean value in class k for the j^{th} feature and x_{kji} is the i^{th} observed value in class k for the j^{th} feature. The FS can be seen to be a ratio of the inter- versus intra-class variance of a feature. A feature that has large separation between the classes and small variance within each class will have the highest FS. Because the Fisher Score assesses the importance of a feature by the ratio of the inter- versus intra-class variance it may unnecessarily bias the selection of features with extreme inter- or intra-class variance but not features with excellent predictive value that have the combination of large inter-class variance and small intra-class variance. Proposed in this thesis is the novel use of Pareto Fronts in §5.6.4 that can potentially overcome such problems.

5.6.2 Classification with in-built feature selection

As outlined previously in §5.4.1 and §5.4.2, PFDA and RDA have in-built feature selection. The classification models themselves, alter the weighting or even inclusion of features in the model. This information can be extracted after the PFDA or RDA model is optimised for classification. How this form of classification performs, measured by classification error of the test data, is addressed in Chapter 6.

5.6.3 RandomForests

RFs can easily incorporate feature importance ranking methods because each tree of the forest only uses a subset of the available training data in its construction. Outlined here is one of the highly favoured methods of feature selection, that uses the 'out-of-bag' (OOB, not sampled in the tree construction) observations and the change in predictive error when the peak expression values of a feature are randomly permuted.

For each of the N_T trees in a RF there are n_{OOB} out-of-bag observations for each tree. These observations can be run through their respective tree to get a predictive error, which is aggregated across all trees. To test the j^{th} feature's importance, the values of the j^{th} feature are permuted in the OOB data. Now when the altered OOB data are run through their respective trees again, the change in predictive error is an indication of the feature's importance in prediction. The relative increase in error (with respect to other features) from the original OOB error is an indication of the feature's importance to correct classification. Figure 5.6 demonstrates how the predictive error can potentially be altered by permuting one of the features.

5.6.4 Pareto Fronts

Pareto Fronts (PFs) have been proposed as a microarray gene ranking tool (Fleury et al., 2002; Hero and Fleury, 2004) as well as a solution and objective checking method for finding peaks in mass spectrometry (Armananzas et al., 2011). Proposed here will be a modified, novel use of PFs to rank MALDI/SELDI TOF-MS peaks for

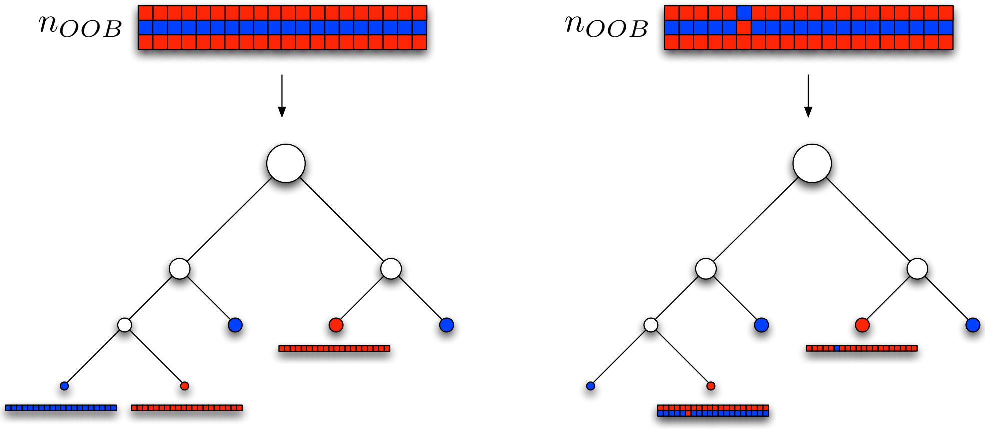


Figure 5.6: Feature importance calculation example for a single RandomForest tree. By permuting the information contained in the sixth feature, the classification of the non-permuted OOB data (left) can be compared to the permuted OOB data (right). The example shows the second observation is correctly predicted at the most left node for the non-permuted OOB data but when the sixth feature is permuted, the loss of information causes the second observation to be incorrectly predicted at the node second from the left.

ability to discriminate between classes. Before this is embarked on, some definitions are required.

Definition 5.1: Dominated feature (Hero and Fleury, 2004). *Given a set of $\{1, 2, \ldots, P\}$ features and ω metrics $\xi_1, \xi_2, \ldots, \xi_\omega$ of interest to be maximised, a dominated feature j is a feature for which at least one ψ,*

$$\xi_\psi(j) < \xi_\psi(q) \;\; \textit{and} \;\; \xi_\zeta(j) \leq \xi_\zeta(q)$$

for $q \neq j$ and $\psi \neq \zeta \in \{1, 2, \ldots, \omega\}$.

The metrics, ξ_ψ for $\psi = 1, 2, \ldots, \omega$, in Definition 5.1 are pre-defined functions that each measure a desirable property of features; the greater the resulting scalar value from the metric for a feature, the more favourable the feature is over others. An example of a metric is the maximum pairwise distance between class means for a feature. If, for example, a metric ξ_ψ for $\psi \in \{1, 2, \ldots, \omega\}$, is optimised by minimisation opposed to maximisation, the situation is easily overcome to fit the original definition by redefining the metric to return the negative of the original value. Thus, by using a definition where larger values returned by metrics for a feature are more favourable, there is no loss of generality. An implementation of this definition as C-code for a set of features is provided in Appendix A.6.

Definition 5.2: Non-dominated feature (Fleury et al., 2002). *A feature that is not dominated as per Definition 5.1.*

The definition of a non-dominated feature is most simply defined as in Definition 5.2 because identification of a non-dominated feature is most economically computed by checking whether it is dominated; if it is not, it is a non-dominated feature.

To illustrate how a non-dominated feature could be identified in a more intuitive but less efficient way, consider proteomic MS peak data for the purpose of finding features that best discriminate between classes. Let intra-class and inter-class variance be the two metrics of interest, labelled ξ_1 and ξ_2, respectively. In this scenario, ξ_1 is optimised by minimisation and ξ_2 is optimised by maximisation with respect to class differentiation. A non-dominated feature is a feature that has the largest ξ_2 value of all the features with its ξ_1 value or less. This can equivalently be defined as a feature that has the smallest ξ_1 value of all the features with its ξ_2 value or greater.

Definition 5.3: Pareto Front (Fleury et al., 2002). *The set of all features that are non-dominated as per Definition 5.2. i.e. The set of features:*

$$\{j' \in \{1, 2, \ldots, P\} \mid j' \textit{ is non-dominated } \}.$$

The Pareto Front (Definition 5.3) is the set of features that form an outer 'shell' around the other features in the ω-dimensional metric-space. As discussed in Hero and Fleury (2004), Pareto Fronts are a way of determining important variables without defining an objective function that returns only a single scalar value. As such, no one feature may be the optimal solution (in this context, most discriminatory peak), but a set of features will be considered equally important.

A Pareto Front (PF) is illustrated by a mock example in Figure 5.7 with two criteria ξ_1, ξ_2 to be maximised. Here, not only is a Pareto Front shown (PF 1) but subsequent Pareto Fronts (PF 2, 3 and 4) by applying the same algorithm for determining non-dominated features once features from the previous PF are removed. This can of course be repeated until all features are classified as belonging to a particular Front. By determining what Pareto Front a feature belongs to, a ranking of the feature's relative importance is created.

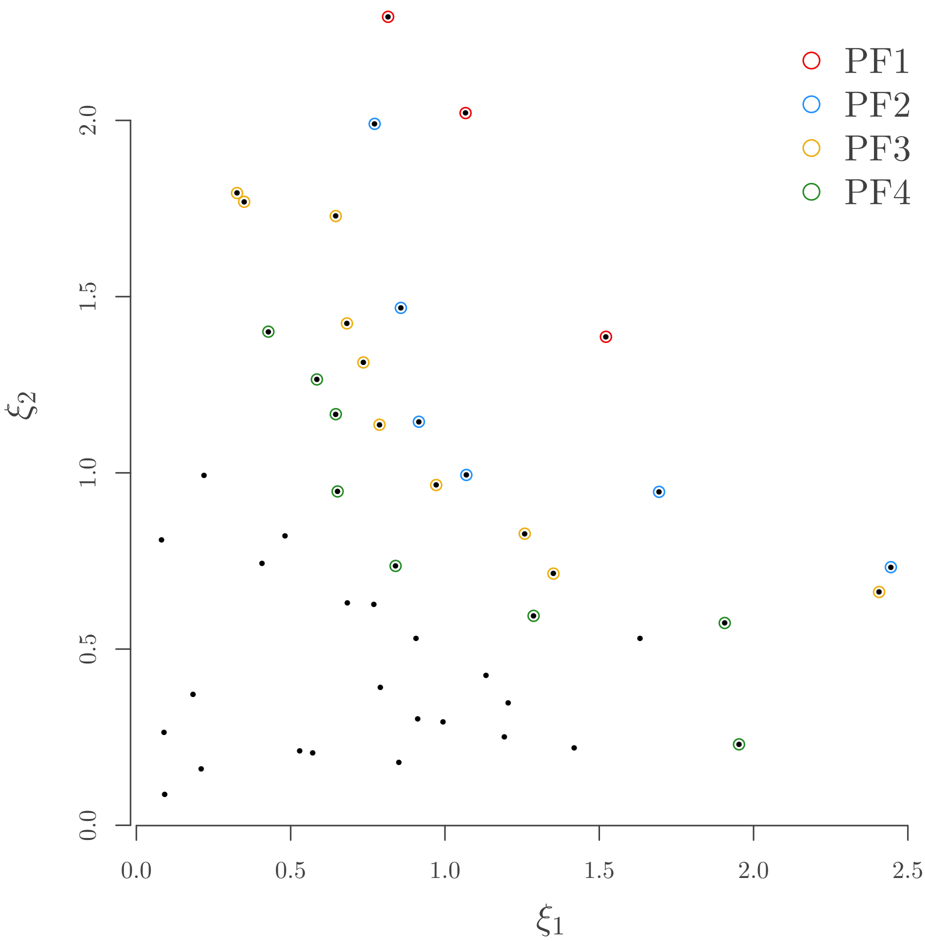


Figure 5.7: An example with two criteria ξ_1, ξ_2 to be maximised and the first four Pareto Fronts.

In the context of repeated measures and time-dependent microarray data, multiple observations are available to estimate the important features. The formula suggested by Fleury et al. (2002) to order features (genes) by importance to the outcome is

$$RF_f\,(j) = \frac{1}{\prod_{t=1}^{T} M_t} \sum_{m_1=1}^{M_1} \sum_{m_2=1}^{M_2} \cdots \sum_{m_T=1}^{M_T} \Delta_f^{(-m_1,-m_2,\ldots,-m_T)}\,(j)\,,$$

where there are $t = 1, 2, \ldots, T$ time points in which expression is measured and M_t samples per time point, and $\Delta_f^{(-m_1,-m_2,\ldots,-m_T)}\,(j)$ is the indicator function of gene j being on the first f Pareto Fronts with samples $m_1, m_2, \ldots, m_T$ removed. This is a leave-one-out estimator.

Such a metric is not applicable for the TOF-MS data analysed in this thesis as they are not time-course data. The proposal here is to rank MS peaks by importance using a cross-validated PF metric based on G-folds as opposed to leave-one-out, for stability and re-sampling of the data for precision (using similar principles as discussed in unbiased estimation, §5.2). Define the discriminatory strength w of feature j as

$$w\left(j\right) = \frac{1}{RG}\sum_{r=1}^{R}\sum_{g=1}^{G}\max_{f\in\{1,2,\ldots\}}\frac{\Delta_f^{(-\boldsymbol{x}_{rg})}\left(j\right)}{h\left(f\right)}, \tag{5.6}$$

where R is the number of re-samples or reshuffles of the data, G is the number of folds in the cross-validation, $\Delta_f^{(-\boldsymbol{x}_{rg})}\left(j\right)$ is defined as an indicator of being a member of the Pareto Front f with the removal of samples in fold g and re-sampling r. The function h is a function to weight Front membership. The function $h(x) = x$ provided sensible results.

From Equation (5.6) it can be seen that PFs provide an alternate method of feature ranking; a feature's importance is relative to other features. Because features are assessed in comparison to each other, implementation of Equation (5.6) can actually create the PF ranking of all features simultaneously for each reshuffle or cross-validation. Such an implementation is provided in Appendix A.6.

An example for a peak j using only a $R = 1$ sample of the data and 2-fold ($G = 2$) cross-validation, where peak j is in the second PF in fold one and the third PF in fold two, would result in

$$\begin{aligned} w\left(j\right) &= \frac{1}{2}\left(\max\left\{\frac{0}{1},\frac{1}{2},\frac{1}{3},\frac{1}{4},\ldots\right\} + \max\left\{\frac{0}{1},\frac{0}{2},\frac{1}{3},\frac{1}{4},\ldots\right\}\right) \\ &= \frac{1}{2}\left(\frac{1}{2}+\frac{1}{3}\right) \\ &= \frac{5}{12}. \end{aligned}$$

As seen above, using $h(x) = x$ provides the 'average' of the inverse PF the feature lies on for each fold to provide a rank of feature j based on the weight calculated between 0 and 1.

Objective functions

An advantage of PF feature ranking is it can be customised to simultaneously assess as many objectives as deemed appropriate for the application. In this content it is of importance to find features that provide information to successfully differentiate between classes. Note that a feature that differentiates between two classes but not the $K-2$ remaining classes has important utility as it may be combined with other features that can help distinguish the remaining classes. Such features would be likely to be overlooked by FSs.

Used here are three criteria that are to be optimised:

(1) ξ_1, minimum intra-class variance (to minimise).

(2) ξ_2, inter-class variance (to maximise).

(3) ξ_3, maximum inter-class distance (to maximise).

Each of the criteria listed above would be indicative of an important discriminatory feature. A small value of ξ_1 suggests a feature has a class with observations closely located to each other. Criteria ξ_2 and ξ_3 are similar in the sense they are both optimal when maximised, but a large ξ_2 indicates a feature where the K classes have good separation relative to other features and large ξ_3 indicates there are at least two classes that have large mean separation.

Comparing Pareto Front feature ranking to other methods

To compare PF feature ranking with FS and RF feature ranking, a dataset of random Gaussian observations was generated. The dataset consisted of $P = 1000$ features and $K = 3$ classes with sample sizes $n_1 = 60$, $n_2 = 50$ and $n_3 = 40$. Only 25 of the $P = 1000$ features were generated where the classes had different true mean values (i.e. were differentially expressed). The mean of the Gaussian observations for all classes were 0 for the first 975 features and the observations in the remaining 25 features were from normal distributions where the mean of each class was determined by a single random observation from a uniform $(-0.5, 0.5)$ distribution. All standard deviations in each class within each feature were from a uniform $(0.5, 1.5)$ distribution. Figure 5.8 shows PF feature ranking is as good, if not better, than RF feature selection. As expected, FS feature ranking was inferior in performance in selecting truly differentiated features.

From preliminary testing of Pareto Fronts as a feature selection method, it should also be noted a slight deviation from Definition 5.1 was implemented that achieved superior results. Instead of a PF being defined as the PF in all dimensions (the $\omega = 3$ criteria), the PF included any feature that was on a PF in two of the dimensions (visualised as the projection of the criteria values onto the plane of two criteria, of which there are $\binom{3}{2} = 3$ pairwise projections). The looser criteria allow more flexibility as optimisation of two criteria is sufficient evidence to suggest a feature is important.

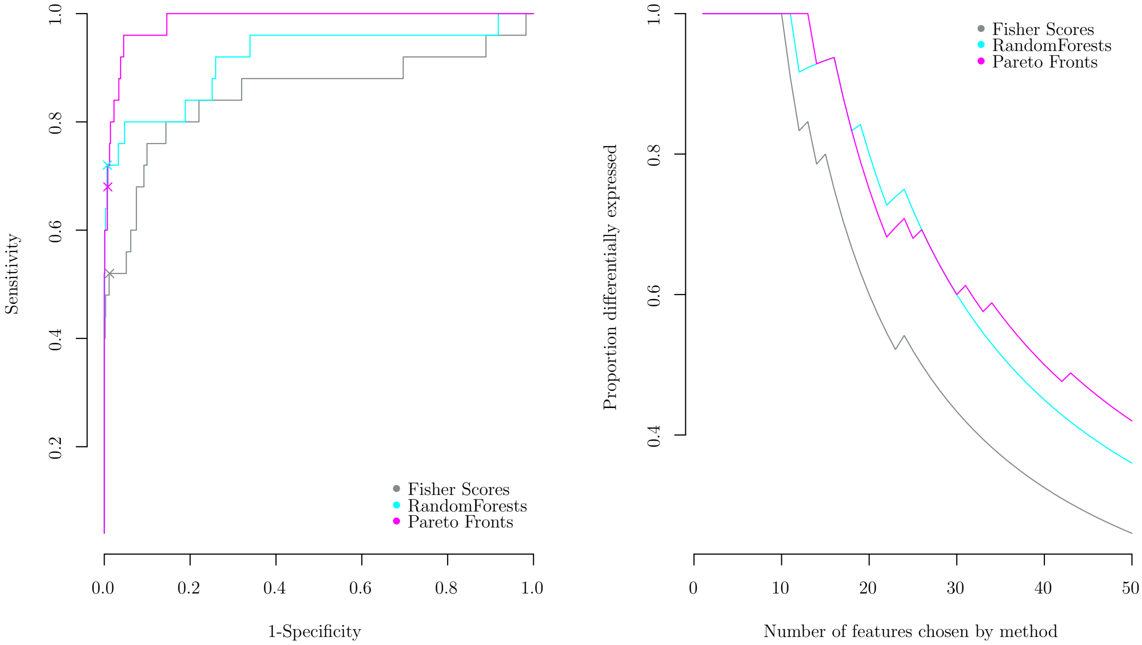


Figure 5.8: (a) A receiver operator curve showing the sensitivity and specificity of the feature ranking methods in selecting the 25 differentially expressed features, and (b) the cumulative proportion of truly differentiated features selected by the feature ranking methods for the number of selected features by the methods ranging from 1 to 50.

5.7 Summary

A brief summary of the discrimination methods, both statistical and computational, can be found in Table 5.2 for easy reference. The attributes listed are discussed in their relevant sections but Table 5.2 provides a comparison of these attributes with the other methods. The attribute ‘proteomic predictive qualities’ is based on results that will be shown in the next chapter, Chapter 6.

The three feature-ranking methods: FSs, RFs and PFs offer very different approaches to feature ranking. FSs are a traditional method of selecting features that differ in means based on statistical significance. This method ignores the potential correlation between features and is likely to over emphasise the importance of features that have very large inter-class variance or very small intra-class variance. RFs evaluate a feature’s importance by effectively removing the information contained within that feature in the OOB data, via random permutation, and determining the subsequent change in classification error. RFs give a good indication of not only the discriminatory power of features but their performance when taking into account other features available. PF ranking is novel to MS peak ranking and the PF implementation explained here is an extension of previous PF methods. PF ranking can be thought

Table 5.2: Summary of attributes of discrimination methods used.

	Model					
Attribute	LDA	QDA	PFDA	RDA	RF	SVM
In-built feature selection	🔴	🔴	🔵	🔵	🔵	🔴
Invariant to transformation	🔴	🔴	🔴	🔴	🔵	🔴
Handling of SnLp problems	🔴	🔴	🔵	🔵	🔵	⚪
Model parameter optimisation	🔵	🔵	⚪	⚪	⚪	🔴
Computational time	🔵	🔵	⚪	⚪	🔴	⚪
Handling of multi-class data	🔵	🔵	🔵	🔵	🔵	⚪
Implementation available (R)	🔵	🔵	🔴	🔵	🔵	🔵
Correlated data	🔵	🔵	🔴	🔵	⚪	🔵
Proteomic predictive qualities	🔵	⚪	🔴	🔴	⚪	🔵

Legend: 🔵 good ⚪ average 🔴 poor

of as using methods from both FSs and RFs. PFs incorporate information that is used in FSs but with discrimination of the classes in mind as opposed to statistical difference in class means. Also, PFs are discrimination-focused and features are assessed relative to others such as in RFs, except discriminatory strength is measured by criteria specified *a priori*.

Chapter 6 will assess the strength of the methods discussed in this chapter on the datasets outlined in this thesis. From the results, the feasibility and reliability of a diagnostic tool using current MALDI/SELDI TOF-MS technology on serum and data analysis tools can be assessed.

Chapter 6

Classification results

This chapter presents the application of the statistical and computational methods described in Chapter 5. There are many practical considerations, both at the data creation level and at the level of the parameters involved in classification. Peak expression datasets have so far been characterised by log_2 (maximum) peak intensities but other peak quantification methods are also explored. As discussed previously, MALDI/SELDI TOF-MS peak expression data contain missingness, an issue that is seldom addressed in the literature. The dataset of most interest in this thesis, the GC mice dataset, has its own experimental design issues and these are addressed to obtain optimal discrimination. All analyses are assessed for optimality using discrimination rule success as the quantitative benchmark.

6.1 Towards optimal discrimination

Best practice for spectra classification of MALDI/SELDI TOF-MS data is not yet well established (Wu et al., 2003; Johansson and Ringner, 2004; Meding et al., 2012). It is hoped the following chapter addresses some important issues which are often overlooked in classification. To formulate optimal data handling and analysis procedures, the following must be considered.

- A method to impute missing values. Here, five methods to impute missing data are studied; assignment of the median peak value, assignment of the minimum peak value, random normal values based on the sample standard deviation, imputed values using `missMDA` and `impute` `R`-packages of §4.1.1.
- Data transformation. Many of the discrimination models have distributional assumptions. Here, standard $\log_2$ peak expression is compared to PCA transformed data.
- Peak quantification. Peak expression has been assumed to be maximum peak height to this point. The three peak quantification methods of peak intensity, empirical peak area and Gaussian modelled peak area (outlined in §3.2) are compared.
- The feature selection methods discussed in §5.6. The three feature ranking and selection methods explored here are Fisher Scores (FS), Pareto Fronts (PF) and RandomForest (RF) variable ranking.
- Classification methods. The six methods compared for effectiveness are: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), pairwise fusion discriminant analysis (PFDA), regularised discriminant analysis (RDA), RandomForests (RF) and support vector machines (SVM).
- Spectra pre-processing normalisation methods from §3.1. Empirical quantile normalisation (EQN) is compared to the standard total ion current normalisation (TCN).

The above, which will be referred to herein as expression construction and analysis elements, is not an exhaustive list of issues to be handled in producing spectra classification. However, should each possible combination of the expression construction and analysis elements be explored, $5 \times 2 \times 3 \times 3 \times 6 \times 2 = 1080$ discrimination models would be required to be fitted. Because of the computational time required to optimise and evaluate the models using the cross-validation process outlined in Figure 5.2, an exhaustive analysis examining all the combinations of the expression construction and analysis elements is infeasible.[1,2] To counter this issue, individual factors will progressively be assessed whilst holding all others constant. This of course does not equate to an optimised process but should indicate methods most likely to be successful for other datasets.

This chapter is organised as follows. The standard classification approach presented in Chapter 5 is elaborated on and refined. Then, important GC mice data-specific considerations are assessed. These are in addition to the expression construction and analysis elements identified that are applicable generally to MALDI/SELDI TOF-MS data. The replicates available in the GC mice data provide an opportunity to determine whether pooling of replicates leads to improved classification. The effect of replication on the success of a classification model will provide important insight into future experimental designs. Sections that investigate the core expression construction and analysis elements above are then finally optimised step-wise.

6.1.1 A standardised approach

The (random) allocation of training and test data as well as GFCV model and parameter optimisation is a standardised process throughout, as per Figure 5.2, for each differing set of expression construction and analysis elements. An optimal predictive model, within the given set of expression construction and analysis elements, is created 100 times on randomly selected training data to demonstrate the stability of the model's predictive utility on test data. Then, 100 error proportions on the training complement (test dataset) are observed for central tendency and variability in assessing the best method. Although the 100 error proportions estimated from the test datasets will not be independent of each other, as the test data will have common spectra from one sampling to another, the test data will be *independent of the training data* in every case. There is not a specific reason 100 repetitions of this process was chosen other than it provided enough error proportions to provide insight of distribution and variation that occurs with these outcome values. The number of common observations from one sampled test dataset to another is simply a hypergeometric random variable. For clarity, observations in the expression data derived from the identified peaks in spectra used in the classification will be referred to as peak expression vectors. These are the n vectors $\boldsymbol{x}_i$, $i \in \{1, 2, \ldots, N\}$ allo-

[1] To run the process outlined in Figure 5.2 on a MacBook Pro7 (Intel Core 2 Duo 2.4 GHz, 3 MB L2 Cache, 8 GB Memory) under one set of constraints takes on average more than 24 hours to run (run 100 times on different training/test data splits to be described later). This average is largely driven by the most computationally intensive methods in RF and PF.

[2] Other spectra pre-processing methods would benefit from inclusion in the list of factors to optimise, if not to provide quantitative merits between competing methods, to determine their overall effect on classification. However; smoothing, baseline subtraction and peak alignment were all able to be evaluated by inspection while the effects of normalisation methods are less straight forward. Because of the computational limitations, normalisation was the focus here.

cated to the training data and the remaining $N-n$ vectors $\boldsymbol{x}_{i'}$, $i \neq i' \in \{1, 2, \ldots, N\}$ allocated to the test data.

Since the aim is to find an optimal combination of expression construction and analysis elements for classification, the training and test data were kept completely separate. To estimate and assess an unbiased predictive error in a diagnostic setting, the test data should not be used to inform the model generation at any point. Furthermore, the GC dataset has the additional complication of replicate spectra per mouse. To avoid downwards-biased error, replicate peak expression vectors for each mouse were all (randomly) placed either in the training or test datasets but importantly do not span both. Placing replicates across the training and test datasets would provide information about the test dataset in the training dataset as it is a reasonable assumption that peak expression vectors from the same mouse will be correlated. This correlation was shown in the empirical pairwise correlation matrix heatmap in Figure 4.8(a).

This chapter is largely descriptive in nature, although the results are compared using the quantitative outcome of prediction error on the test data. The term descriptive is used since the recommendations are based on tables and figures of the data produced. The ability to make formal statistical inference beyond this should be made with trepidation (Dietterich, 1998); the 100 final classification models optimised on the training data are highly correlated and the predictive error estimates on the 100 test datasets are dependent, because of the training and test data commonality from re-sampling for each of the 100 final classification models. However, the method of re-sampling offers excellent insight into the stability of the models produced and will help to select the appropriate expression construction and analysis elements for classification.

There are data treatment procedures constant throughout the analysis. The training/test data split was made using a random allocation of $\frac{2}{3}$ to the training data and the remaining $\frac{1}{3}$ to the test data, as discussed in Chapter 5. The training/test data allocation was performed 100 times at random to create 100 final models that generate the 100 predictive test data errors for each set of expression construction and analysis elements. Five-folds were used in the cross-validation on the training dataset. This allowed a consistent approach for all datasets, importantly allowing a sufficient number of observations available in the training data to have appropriately representative folds of all the classification groups, especially in the multi-class case. As per the iterative method in optimising each model (Figure 5.2), different-sized sets of the highest ranked features are used. The number of feature sets considered for each iteration was kept to 20 per fold in each iteration. The first feature set in each iteration was chosen to be the top two ranked peaks and the last feature set was the entire set of peaks.[3] By limiting the number of feature sets to 20, the computational time was reduced without much loss in predictive utility (from preliminary testing). In the case where two or more models in the iterative optimisation process on the training data produced the (same) lowest GFCV error, the model using the smallest number of features was algorithmically preferred.

To further simplify the process, initially, all five datasets were modelled for an outcome with only two outcome classes (female or male for the asthma datasets and a dichotomised outcome of disease or control for the remaining datasets). Classification was undertaken on all available class information in the final analysis if the dataset contained greater class granularity. For example the GC mice dataset has five outcome classes that are assessed in the final analysis: WT, IL6, FFStat3, FFIL6 and FF. This provided justification for choosing five-fold cross-validation: it allowed like-with-like comparisons as the five-fold approach was kept constant over all models.

The discrimination results in this chapter are represented as multiple line bar graphs to compare different expression construction and analysis elements on the error proportion. The bars for particular error values in the plots have heights relative to the proportion of the 100 error estimates. There is an associated orange line that depicts the range of the errors and a maroon bar under the histogram error lines showing the 75th, 50th and 25th percentiles (reading left-to-right with worsening prediction and thus larger error). This method of visualisation allows more detailed assessment of error estimates than a traditional histogram and more exact information than a traditional boxplot. It also allows depiction of the discrete nature of the error estimates as they can take the possible values of $0, \frac{1}{N-n}, \frac{2}{N-n}, \ldots, 1$, where the number of allocated test observations, $N-n$, is different for the five datasets used. In addition to these plots, the expression construction and analysis elements involved in the discrimination in each case are listed below the plots.

The starting set of expression construction and analysis elements were chosen to be:

- Missing values: imputed by the median value of the peak expressions observed for that peak.
- Data transform: $\log_2$.
- Peak expression: maximum peak height.
- Feature selection: PF.

[3] This upper limit was reduced to the maximum number of features the discrimination model was able to use with success. For example, parameter estimates are not possible when $n-K \geq p$ for LDA (§5.3.1). The feature sets considered in optimisation were then the 20 feature sets starting with the feature set containing two features up to the final feature set containing the new maximum number of features.

- Classification method: SVM.
- Pre-processing normalisation: EQN.

Note that expressions for each peak were centred and scaled by the observed standard deviation since classification methods such as SVM are sensitive to expression heterogeneity between features (scaling is actually performed internally by default in `LIBSVM`). Specific model parameters such as γ and c of the SVM model were optimised as additional parameters in the five-fold cross-validation (Figure 5.2). It is worth noting that the median peak expression was chosen as the initial imputation method because of its simplicity and its higher likelihood to conform to the assumption of normality within peak expressions, implicit in the PCA transform (Fang and Han, 2013) considered early in the process. As a basic imputation method, it has a desirable property that it will add no information to differential expression between classes within peaks because of its centrality. Conversely, it does not utilise the information potentially available from missing values which may be a result of expressions below the detection threshold, as discussed previously in §4.2.

6.2 Optimising discrimination of the GC mice dataset

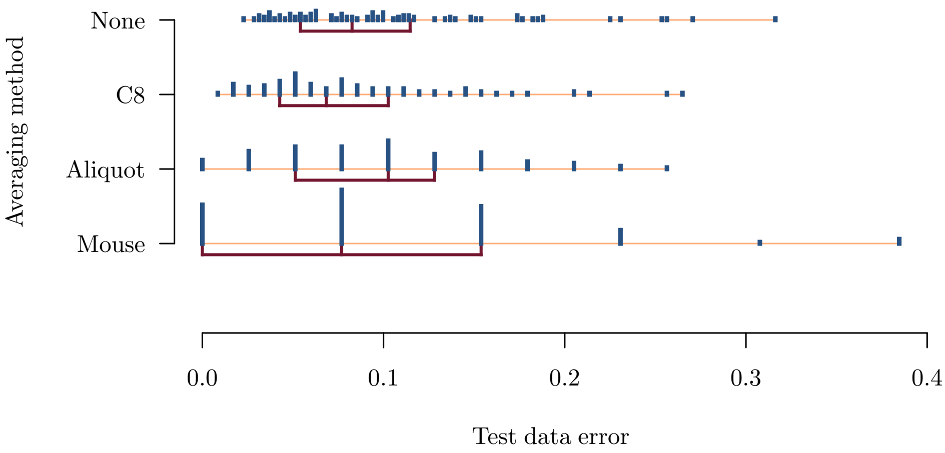
Because of its multi-chip and replicate design (Figure 4.6), the GC dataset provides an excellent opportunity to explore factors that affect classification signal beyond the expression construction and analysis elements discussed so far. This section concerns itself with GC mice specific nuances prior to the optimisation of the expression construction and analysis elements on all five datasets.

The effect on the discrimination of the GC mice data by varying the amount of replication for training the classification model and the amount of averaging peak expression vectors is investigated. Additionally, the accuracy of class prediction for individual mice is presented to establish whether there exists a relationship between the classification model's certainty and correct prediction. The outcomes of these GC mice specific considerations may provide important information about the best experimental designs and diagnostic rules for future studies.

6.2.1 Averaging over peak expression vectors

The GC mice data contain 27 replicate peak expression vectors per mouse, nine from each of the three chips as well as three within each of the nine C8 fractionisations. It is not known what the effect of averaging peak expression vectors has on the classification signal, and more importantly, what level of averaging within experimental subjects (mouse, aliquot, C8 fractionation or none) yields the best classification.

Figure 6.1 demonstrates the proportion of incorrectly predicted classes in the test data peak expression vectors to GC and control groups under different peak expression vector averaging schemes. The height of the vertical lines represent the frequency a particular error proportion on the test data was observed. The number of error proportion values that are possible when peak expression vectors are averaged become smaller, a result of a smaller number of peak expression vectors in the test data, $N - n$. The 'mouse' averaging method denotes the averaging of peak intensities over all 27 replicate peak expression vectors per mouse. Therefore each mouse only has only one peak expression vector for discrimination which reduces the total number expression vectors from 1080 to 40 (a training test dataset split of 27/13). Similarly, the 'aliquot' averaging method takes the mean peak intensities within aliquot for each mouse, resulting in 120 peak expression vectors for classification (81/39 training/test split). 'C8' averaging denotes a dataset with 360 peak expression vectors (243/117 training/test split) with averaging performed on the peak expression vectors of the three technical replicates within each C8 treatment of each mouse. The 'none' method retains all 1080 peak expression vectors (727/351 training/test split).



Spectra normalisation	Classification method	Feature selection	Peak quantification	Data transformation	Imputation method
EQN	SVM	PF	Peak height	$\log_2$	Peak median

Figure 6.1: Discrimination results for GC mice dataset using different levels of averaging of the peak expression vectors.

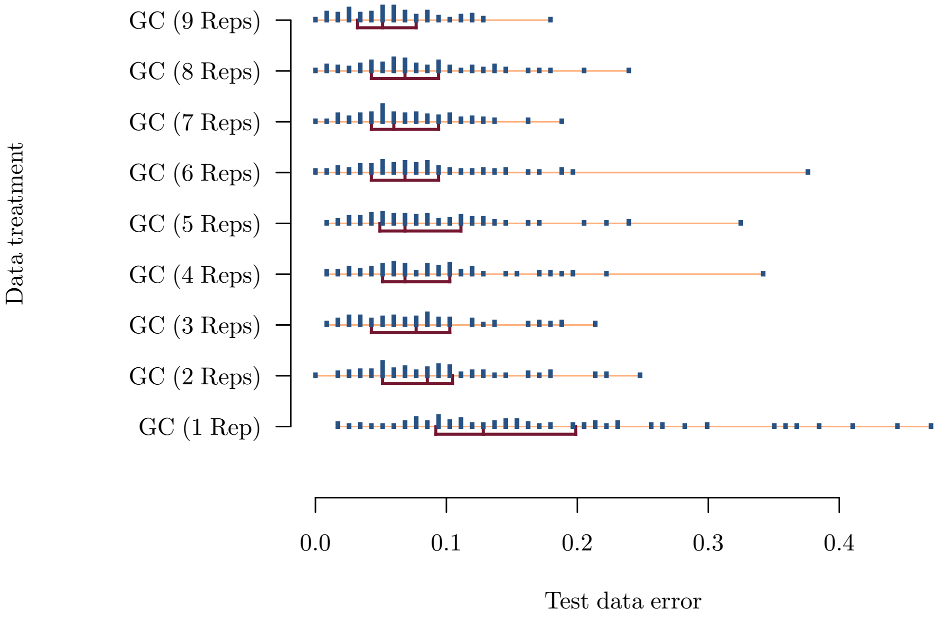
The 'none' and 'mouse' averaging methods provided an increased range and generally larger observed test data error. The 'mouse' averaging method had the greatest number of estimated 0 error proportions but this is a result of the binomial probabilities of a limited number of 'trials' (the test data, $N-n = 13$) for a small underlying probability of misclassification. The 'mouse' method of averaging is inferior to the 'aliquot' and 'C8' methods because of its increased variability, as well as increased 25^{th} percentile and maximum estimated error. The 'C8' averaging method is the preferred method and used hereafter as it has the lowest observed median and 25^{th} percentile (error proportions of 0.068 and 0.103 respectively).

Some averaging of peak expression vectors to produce more favourable results is to be expected because averaging is performed on observed expressions and missing values do not contribute to average values. For example, if spectra a, b and c are to be averaged and the expression for peak p is missing for spectra a and b, the averaged peak expression vector will contain spectrum c's peak expression for peak p. In this way, averaging performs an imputation of sorts by reducing the number of missing values. In the hierarchical modelling in §4.2, the residual error was a large variance component, implying the technical replicates were a large source of expression variability. It is plausible that the increase in information from the pooling of peak expressions and averaging over technical replicates using the C8 averaging scheme offsets the loss in sample size by the decrease in the number peak expression vectors via averaging.

6.2.2 Threshold of replicates to achieve successful discrimination

The GC mice dataset provides an opportunity to assess the effects of subject replicates in the training data on the predictive error in the test data. This was performed by using the 'C8' averaging scheme that was deemed to be most successful, however, the final 100 discrimination models were generated using training data with only $1, 2, \ldots, 8$ or 9 replicates from each mouse. This allowed the detection of any relationship between the number of replicates used in the training data and the predictive ability of the model. To explain this approach, consider the case of three replicates used in the training set to generate the model: 27 mice are randomly allocated to the training set and the remaining 13 mice to the test dataset. Three randomly selected 'C8' averaged peak expression vectors from each of the 27 training data mice are used in the training dataset: resulting in 81 peak expression vectors to use as the training data. Once the model is empirically optimised on the training data, the predicted class of all the available nine peak expression vectors in the remaining 13 training data mice ($9 \times 13 = 117$) is estimated. The test data will contain 117 peak expression vectors whatever the number of replicates in the training data.

Figure 6.2 shows the discrimination errors in the test data using a varying number of ‘C8’ averaged peak expression vectors in the training data. There is a clear advantage in replication as would be expected. The most striking difference is that between one replicate and two. The median test data error proportion was 0.128 for one training replicate and 0.085 for two replicates. The corresponding median error proportion for nine replicates was 0.051. For greater than one replicate, all median test data error proportions were below 0.100 and, with the exception of eight replicates, the median test data error proportion remained the same or decreased as replicates in the training data increased. However, the range of test data error proportions did not uniformly decrease for increasing replicates. Four, five and six replicates had extreme test data error proportions above 0.300 (Figure 6.2). Using a greater number of re-samplings might provide some insight into these effects.



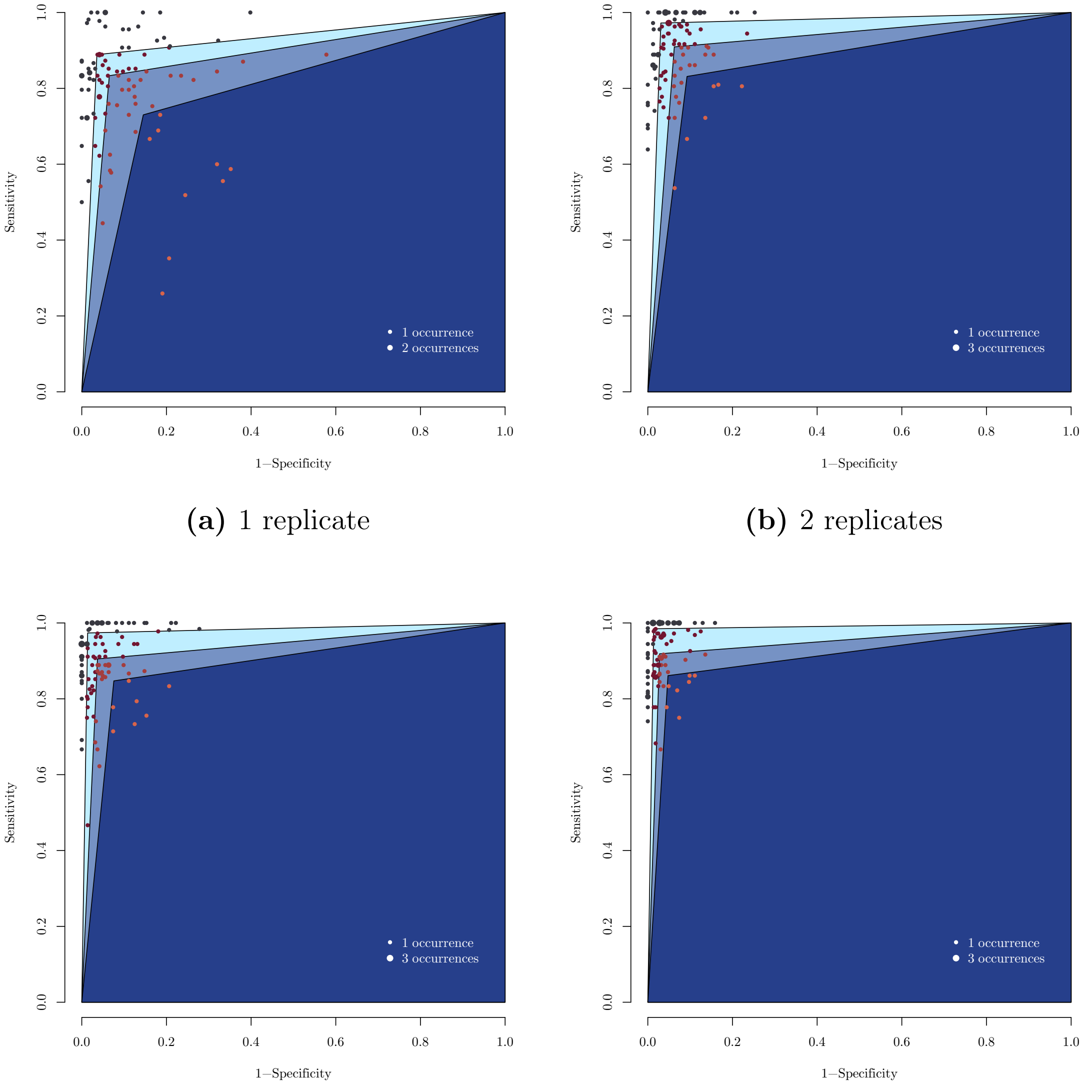
Spectra normalisation	Classification method	Feature selection	Peak quantification	Data transformation	Imputation method
EQN	SVM	PF	Peak height	$\log_2$	Peak median

Figure 6.2: Discrimination results for the GC mice dataset for varying replication in the training data (C8 averaged).

As the median test data error was generally reduced with no outstanding point of improvement for increased replicates (beyond three replicates) used in the model training, the threshold to limit the number of replicates used in future experiments are likely a cost versus benefit decision. That being said, having at least two replicate peak expression vectors in the training data is clearly a large advantage over one. It is established that replication is important in microarray studies (Lee et al., 2000; Yang and Speed, 2002; Allison et al., 2006). Replication has been shown here to be an important consideration for MALDI TOF-MS data. How to handle missing data and whether to average peak expression vectors are additional considerations in the MALDI TOF-MS context.

To further investigate the benefits of increasing replication in the training data, Figure 6.3 shows modified sensitivity and specificity plots to demonstrate the diagnostic utility of the classification models. Sensitivity is the proportion of peak expression vectors from GC mice in the test data correctly predicted as GC mice (true positives) and specificity is the proportion of peak expression vectors from control mice in the test data correctly predicted as control mice (true negatives). The sensitivity and specificity for each of the 100 test data errors are plotted (or plotted as larger circles for repeated sensitivity and specificity observations). The lines separating areas on the plot are the 75th, 50th and 25th percentiles for the observed $(1 - \text{specificity}, \text{sensitivity})$-tuples. From these plots it can be observed that the specificity outperformed the sensitivity in general. This can also be seen in Table 6.1. From this it can be inferred the predictive model was slightly biased towards predicting control status for cancer mice, rather than predicting cancer status for control cases. This artefact is investigated in the next subsection, §6.2.3.

Table 6.1 provides a summary of the median and 25th percentile sensitivity and specificity of the replicate analysis. The 25th percentile is also included as it gives an indication of a more conservative estimate of the true classification sensitivity and specificity, and an indication of the stability of the classification model with regards to the median. The area under the curve (AUC) for a nominated percentile (here: 75th, 50th and 25th) is calculated as the area of the quadrilateral created by the given percentile's $(1 - \text{specificity}, \text{sensitivity})$ coordinate and the $(0, 0)$, $(1, 0)$ and $(1, 1)$ coordinates (this AUC calculation is equal to the average of the sensitivity and specificity). As noted previously, the increase from one replicate to two replicates provided the largest absolute improvement. The median and 25th percentiles generally improve with larger numbers of replicates with the exceptions of four, five and six replicates as noted previously. The nine replicate results yield a impressive 91.9% and 97.2% median sensitivity and specificity, respectively. The 25th percentile of the specificity estimate based on the test data error for nine replicates used in the training data was also remarkably high with 95.2%, however a reduction of more than 5% sensitivity from the median meant the 25th percentile sensitivity was a less promising 86.1%.



(a) 1 replicate

(b) 2 replicates

(c) 3 replicates

(d) 9 replicates

Figure 6.3: Modified sensitivity and specificity plots for replicates on GC mice data.

Table 6.1: The median and 25th percentile sensitivity, specificity and area under the curve for the GC mice data using varying numbers of peak expression vector replicates in the training data (model creation) for test dataset prediction.

	50th percentile			25th percentile		
Replicate(s)	Sens	Spec	AUC	Sens	Spec	AUC
1	0.833	0.935	0.884	0.730	0.855	0.792
2	0.909	0.937	0.923	0.831	0.907	0.869
3	0.905	0.963	0.934	0.847	0.924	0.886
4	0.889	0.970	0.930	0.833	0.944	0.889
5	0.911	0.954	0.933	0.827	0.926	0.877
6	0.889	0.965	0.927	0.822	0.937	0.879
7	0.911	0.967	0.939	0.865	0.926	0.896
8	0.903	0.970	0.937	0.841	0.943	0.892
9	0.919	0.972	0.945	0.861	0.952	0.907

6.2.3 Individual mice and their classification

Of particular interest in the classification models is whether any particular mice or peak expression vectors are consistently being misclassified. Figure 6.4 shows a heatmap of misclassified peak expression vectors using the four original peak expression vector averaging schemes: none, C8, aliquot and mouse. In this figure, the rows represent the 40 mice, where the top 16 rows are the GC mice and the bottom 24 rows are the control mice. Each column for the four sub-plots represents a specific replicate within the averaging scheme. Thus there are 27 columns for the no averaging scheme, nine for the C8 averaging scheme, three for the aliquot averaging scheme and one for the mouse averaging scheme. The colour in the heatmaps depicts the proportion of times the mouse and replicate combination was misclassified when it was allocated to the test data; the darker the colour of the cell the more often the peak expression vector was correctly classified. Note that the ordering of replicates is maintained so any column of a sub-plot corresponds to a triplicate of consecutive columns in the preceding sub-plot. For example, the misclassification proportions of the peak expression vectors in column one of the C8 averaging scheme (Figure 6.4(b)) correspond to the peak expression vectors of first three columns in the no averaging scheme (Figure 6.4(a)).

It can be seen in Figure 6.4 that there are two GC mice (labelled mice 4 and 9) in the GC group that are particularly prone to misclassification across the different averaging schemes. These are the main culprits of the lower sensitivity compared to the specificity; mice 32, 33 and 34 are the largest contributors to misclassification in the control group but to a much lesser extent.

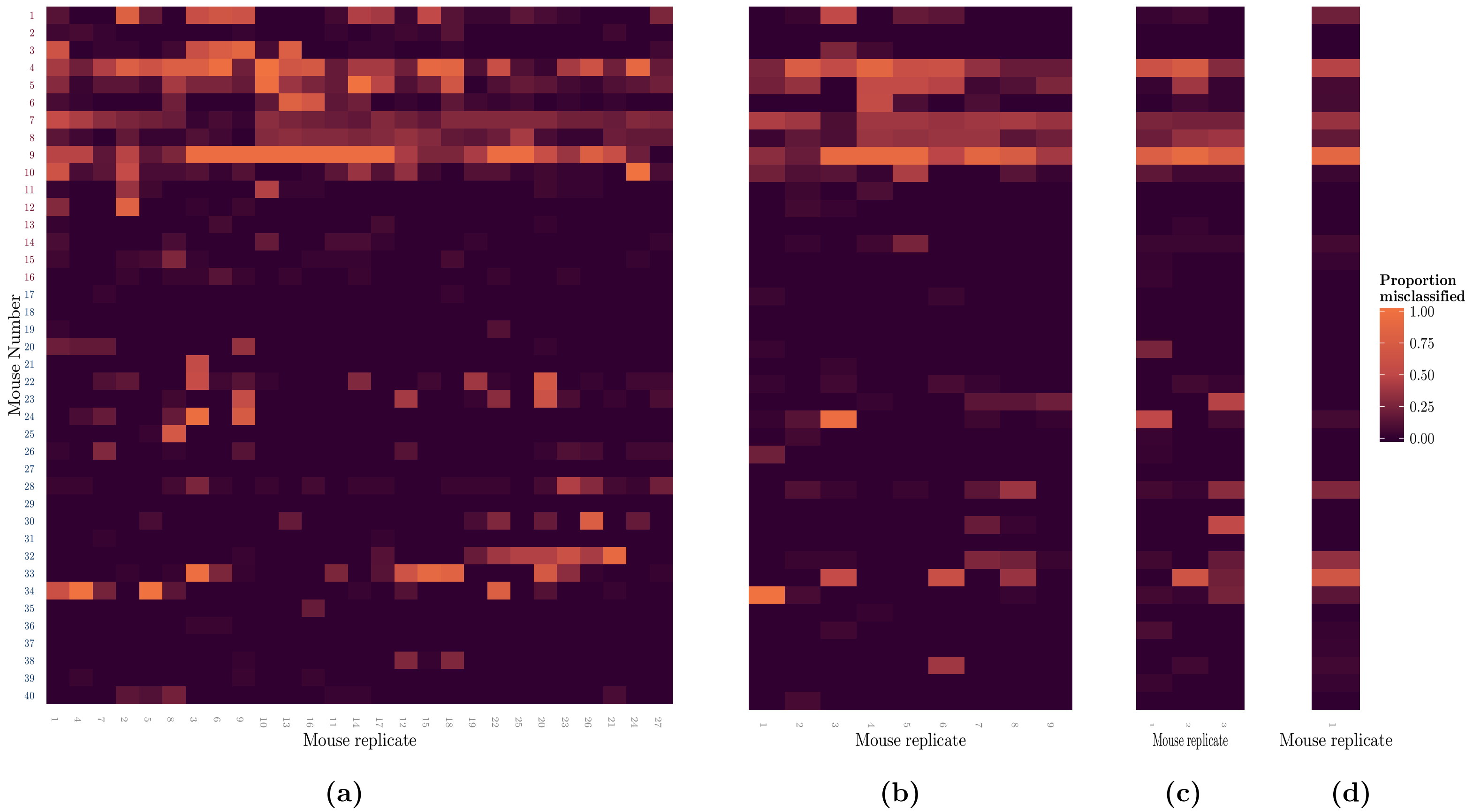


Figure 6.4: Heatmap of the SVM predictive error for GC mice dataset for different peak expression vector averaging schemes: (a) none, (b) C8, (c) aliquot and (d) mouse.

Statistical classification models estimate the probability of observations residing in the K classification groups which then informs the estimated classification. Computational classification models are generally more concerned with the group classification estimate but can also provide probabilities of observations residing in the K groups. Here, SVMs produce probabilities from a logistic distribution using maximum likelihood estimation (Dimitriadou et al., 2011). Examination of these probabilities is of interest, especially for identifying peak expression vectors which have been misclassified most often.

Figure 6.5 shows the average posterior classification probability from the SVM and the proportion of correct classifications of peak expression vector replicates for the C8 averaging method when allocated to the test data. Each colour and shape combination in the two panels (one for each classification group) of the plot represents a different mouse. Each mouse has nine C8 replicates shown on the plot. There is an observable relationship between the average probability and the proportion of correctly classified peak expression vectors. Figure 6.5 shows two solid lines: the vertical line represents the average probability of 0.9 associated with the model estimated classification for a peak expression vector and the horizontal line represents a peak expression vector that has been classified correctly 90% of the time. It is clear from these plots, separated into the two classification groups, that a high average prediction posterior probability (to the right of the vertical line) is associated with successful discrimination on average (above the horizontal line).

Note that the vast majority of points (72%, 259 of 360) in Figure 6.5 lie in the upper quadrant of the plots. That is, peak expression vectors with an average probability of prediction 0.9 or greater and correct prediction 90% or more of the time.[4] Although correct status of the classification will not be known in a diagnostic setting, the probability associated with the estimated classification can be utilised to provide an ‘unknown’ classification group (where the posterior probability is below a certain threshold). This could add utility to a screening test in providing conservative classification by limiting results to those that are most likely correct. For diagnostic tests that return unknown status, a new serum sample could be collected and reanalysed.

[4] In light of LDA models producing superior classification for the GC mice dataset, shown in the following sections, plots similar to Figures 6.4 and 6.5 were produced using LDA classification. As LDA is an inherently probabilistic classifier, the probabilities corresponding to classification estimates might be expected to be more precise. This might indeed be inferred by observation as classifications with an average probability of prediction 0.9 or greater and correct prediction 90% or more of the time increased to 86% (301 of 360) using LDA. To see these figures please refer to Appendix D.

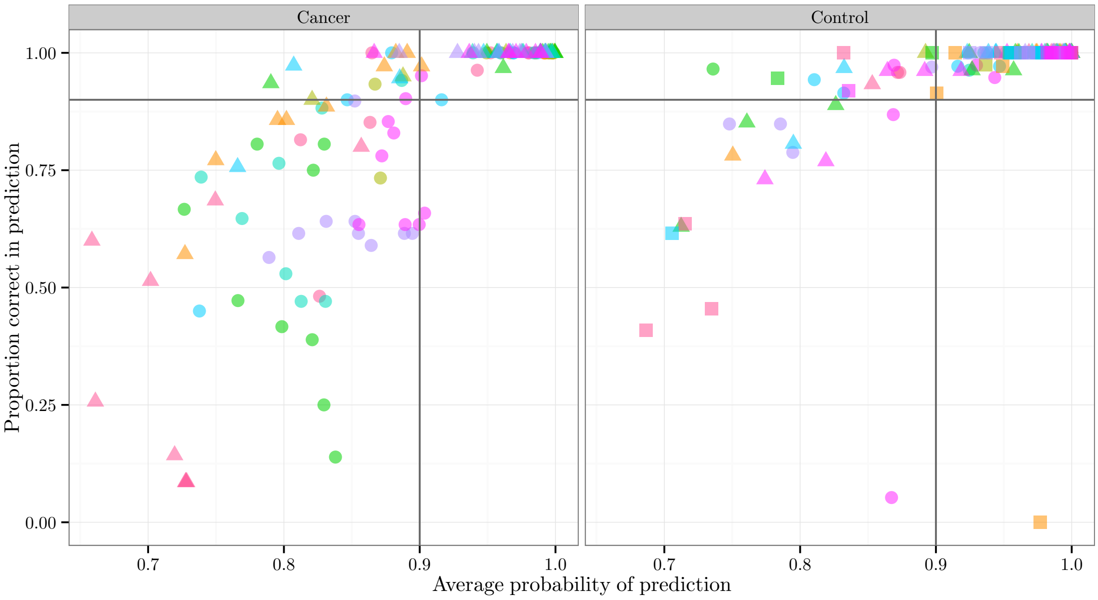


Figure 6.5: Relationship between SVM prediction certainty and prediction outcome for the C8 averaged data. Each distinct colour/character represents the nine replicates of the mice within the GC and control groups. The vertical line represents the average certainty/probability of 0.90 for that particular mouse/replicate and the horizontal line represents the outcome of correct prediction of 90%.

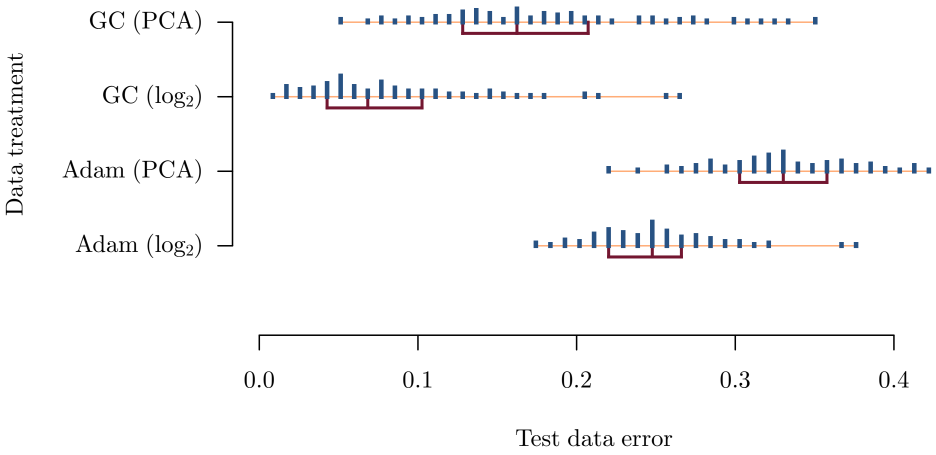
6.3 Comparison of data handling

In addition to the absence of a standardised approach to pre-processing of proteomic MS data, practical data preparation issues of how peak expression is quantified, how missing values are handled and whether dimension reduction techniques should be applied are not standardised. Here, potentially useful methods are compared to provide a preferred basis of analysis.

For this section, the GC mice and Adam et al. (2002) datasets will be used to assess the proposed data preparation methods. The asthma datasets are omitted because of a lack of discriminating signal. The de Noo et al. (2006) dataset is too easily discriminated on a small number of features to have sufficient utility.

6.3.1 Change in basis of data

It is standard to $\log_2$ transform raw expression data to create roughly Gaussian peak expressions within peaks. An additional data preparation method to be considered that may potentially reduce the number of peaks required when optimising the discrimination model, is to PCA transform the $\log_2$ expression data. As discussed in §4.1.1, PCA does not natively handle missing values. To overcome this issue, median peak expression values were imputed to maintain a roughly symmetric distribution of expression within peaks. In doing so, deviation from the assumption of normality in PCA can be minimised. The results of this method versus standard $\log_2$ transformed peak expression data can be seen in Figure 6.6. For both the GC and Adam et al. (2002) data, the PCA transformation of data performed worse and with increased variability. The unambiguous increase in prediction error using PCA transformation of peak expression vectors makes the $\log_2$ transformed expression data the preferred approach.



Spectra normalisation	Classification method	Feature selection	Peak quantification	Data transformation	Imputation method
EQN	SVM	PF	Peak height	⟨variable⟩	Peak median

Figure 6.6: Discrimination results for GC mice and Adam et al. (2002) datasets using varying data transformation methods.

6.3.2 Treatment of missing values

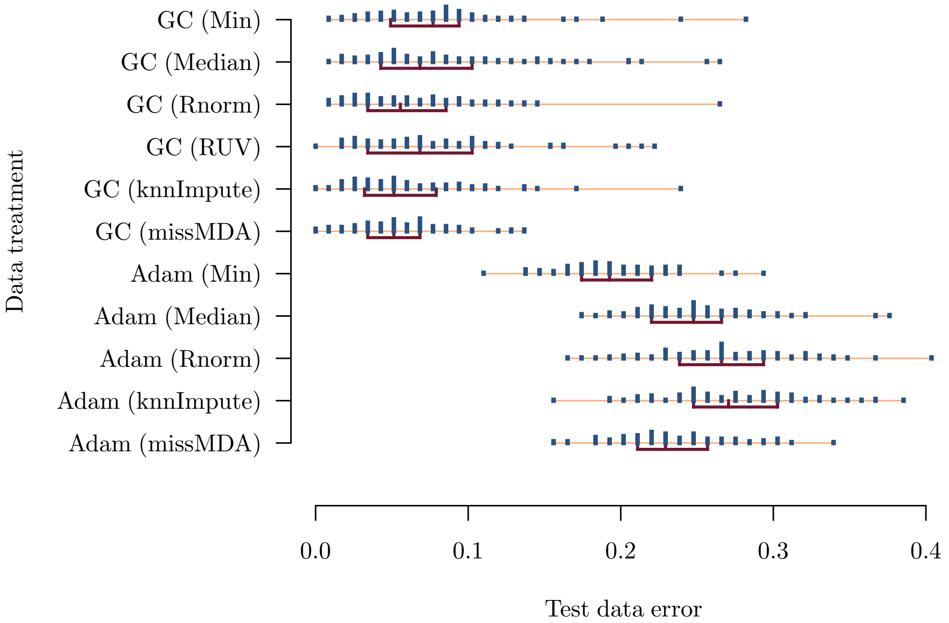
One of the biggest challenges in the analysis of MALDI/SELDI TOF-MS data is the treatment of missing values. So far in this chapter, missing values have been imputed using the median value for the peak expression. Here, four alternatives are assessed: a value less than the minimum observed peak expression for the peak (Min, half a standard deviation below the minimum was used), random normal observations

(Rnorm), imputing values using the `missMDA` package (missMDA, see §4.1.1) and imputing values using the `impute` package (knnImpute, see §4.1.1).

The rationale for each of these methods is as follows:

- Median: does not provide information about up- or down-expression of the peaks in spectra with missing values so is not likely to artificially influence prediction using classification group information. Additionally it will maintain a roughly symmetric distribution of expressions within a peak for discrimination methods that have distributional assumptions.
- Min: uses a value below the smallest expression within the peak. Most missing values are likely to be a result of not detecting a small signal, so it is the most likely value (see §4.2 for justification). The minimum peak expression will hopefully be indicative of the detection threshold. As the data are $\log_2$ transformed, an expression of 0 cannot be log transformed, so an arbitrary distance below (half a standard deviation) the minimum $\log_2$ expression is used.
- Rnorm: random normal observations will not add any classification signal to the data, only noise consistent with the observed data, and the results using this method may additionally show how robust the classification models are to varying values.
- missMDA: imputed values are blinded to class information and therefore do not compromise the error prediction by potentially contaminating the test data with class signal from the training data.
- knnImpute: similarly to missMDA, imputed values are calculated blinded to class information and therefore do not compromise the error prediction by potentially including class signal in the predictive data. This method has been successfully used on microarray data as outlined in §4.1.1.

Figure 6.7 shows the predictive error using the four missing value methods for both the GC mice and Adam et al. (2002) datasets. There was no clear best method for both datasets. The Rnorm and knnImpute methods had the worst outcome for the Adam et al. (2002) data in terms of range of values (stability) and location but performed relatively well on the GC mice data. The Min method was the most successful for the Adam et al. (2002) data with an 25th percentile error better than the median error for all the other missing value methods. The missing value methods for the GC mice data did not vary the estimated test data error greatly. All methods had a median predicted error below 0.100. However the more sophisticated imputation methods of knnImpute and missMDA had superior 25th percentiles. As there was no stand-out method, based on the superiority of the Min method in the Adam et al. (2002) data and the superiority of the missMDA method in the GC mice data, the Min and missMDA methods will be used for the following discrimination analyses.

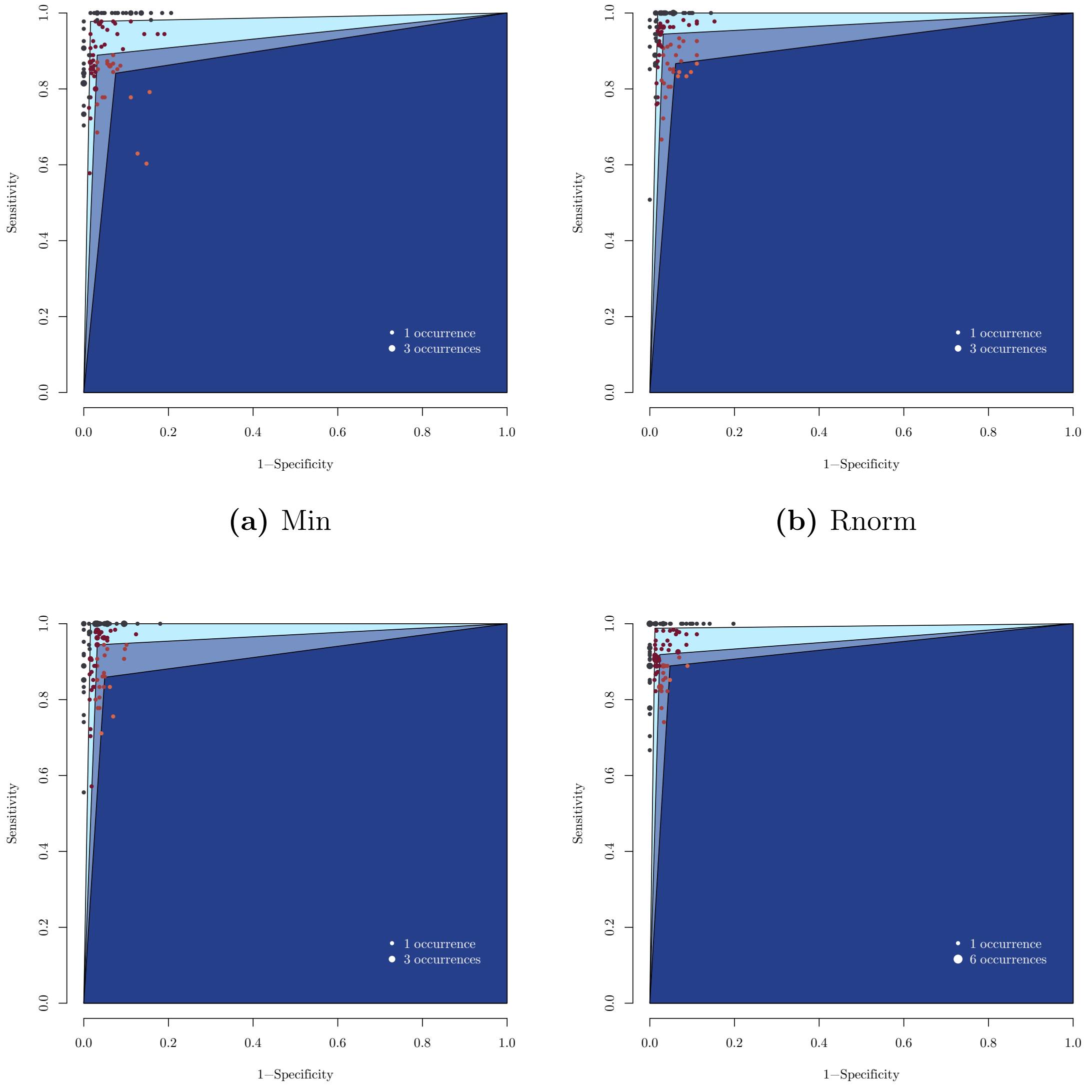


Spectra normalisation	Classification method	Feature selection	Peak quantification	Data transformation	Imputation method
EQN	SVM	PF	Peak height	$\log_2$	⟨variable⟩

Figure 6.7: Discrimination results for GC mice and Adam et al. (2002) datasets using different imputation methods.

The RUV peak expression vector manipulation to average out experimental effects was included in Figure 6.7. It is placed next to the knnImpute method as RUV used knnImpute for missing values. As knnImpute performed better than RUV, RUV adjustment will not be pursued further in classification.

To explore the effects of the missing value imputation method on predictive strength in a diagnostic setting, Figure 6.8 shows the sensitivity and specificity of the four newly introduced imputation methods in classification for each of the predictive test error estimates on the GC mice data. The methods do not appear to change the sensitivity and specificity to a great extent. Like previous sensitivity and specificity plots, there is better specificity on the test data than sensitivity (possibly a result of a larger proportion of control mice than cancer mice used for the training model).



(a) Min

(b) Rnorm

(c) knnImpute

(d) missMDA

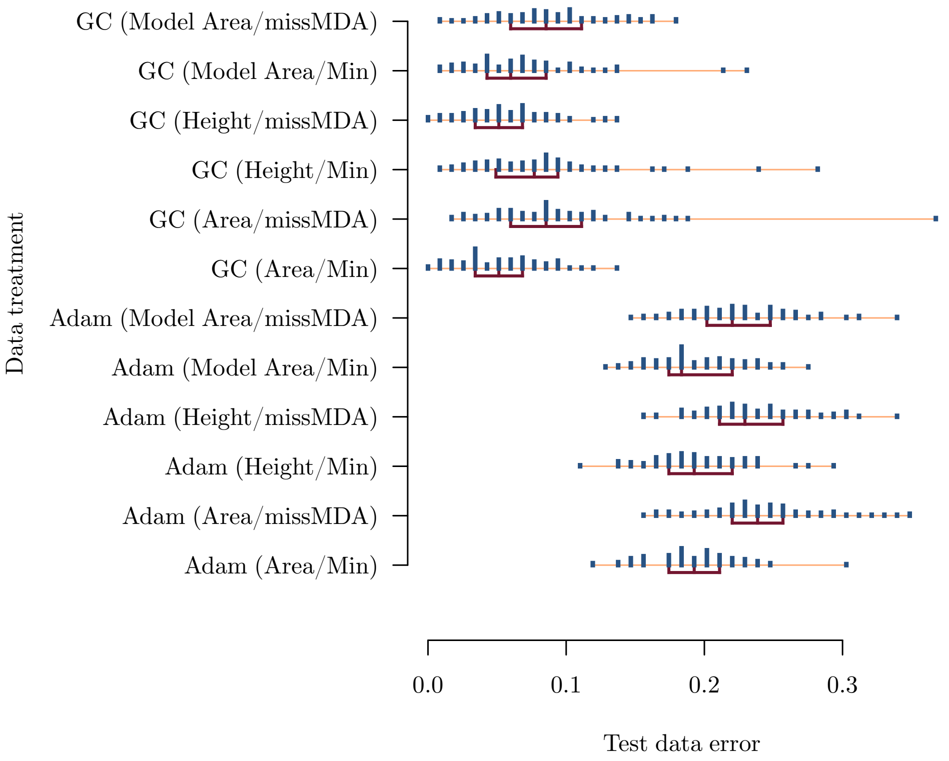
Figure 6.8: Modified sensitivity and specificity plots for different missing value imputation methods on the GC mice data.

6.3.3 Peak expression quantification

Using the two short-listed imputation methods from §6.3.2, an evaluation of the effect of peak quantification on classification is presented here. To this point, peak expression has been assumed to be the maximum peak height measured in the spectra and this is a standard method (Anderle et al., 2004). However, peak expression can also be quantified as the empirical area under the peak (denoted as 'area', the sum of the intensities at m/z-values within the FWHM peak range). The area under a non-linear Gaussian regression on intensities with m/z-values corresponding within the peak FWHM (denoted 'model area', as outlined in the pre-processing section §3.2) is also considered to quantify peptide expression.

Discrimination results using these three peak quantification methods are shown in Figure 6.9 in conjunction with two missing value methods. It is clear that the Min method of imputation provides more accurate prediction, with the exception of the GC mice dataset using peak height and missMDA imputation. The peak height and missMDA combination is the reason missMDA was considered as a favoured imputation method from the previous section, however on the GC mice dataset, the empirical area quantification coupled with Min peak expression imputation provided equally good prediction. On the Adam et al. (2002) dataset, the Min method of imputation was superior irrespective of the peak quantification methods used. The peak quantification methods together with the Min method of imputation performed similarly but the empirical area produced a slightly smaller 25th percentile of test error on the Adam et al. (2002) dataset. The modelled area of both datasets did not perform as well as the empirical area. This may be the influence of peaks for which the iterative non-linear regression algorithm did not converge. This occurred in 2.3% of detected peaks (746 of 31525), in which case the empirical area was used. This may be a source of bias and a disadvantage of the modelled area method on a practical level.

As the Min imputation method provided smaller test data error than missMDA when holding peak quantification constant (with the exception of peak height in the GC mice dataset) and empirical peak area quantification produced the lowest error when paired with the Min imputation method for both datasets, the preferred methods of Min imputation and empirical peak area seem reasonable. This is a surprising result as the maximum peak height is often the assumed peak expression metric (Coombes et al., 2005).



Spectra normalisation	Classification method	Feature selection	Peak quantification	Data transformation	Imputation method
EQN	SVM	PF	⟨variable⟩	$\log_2$	Min/missMDA

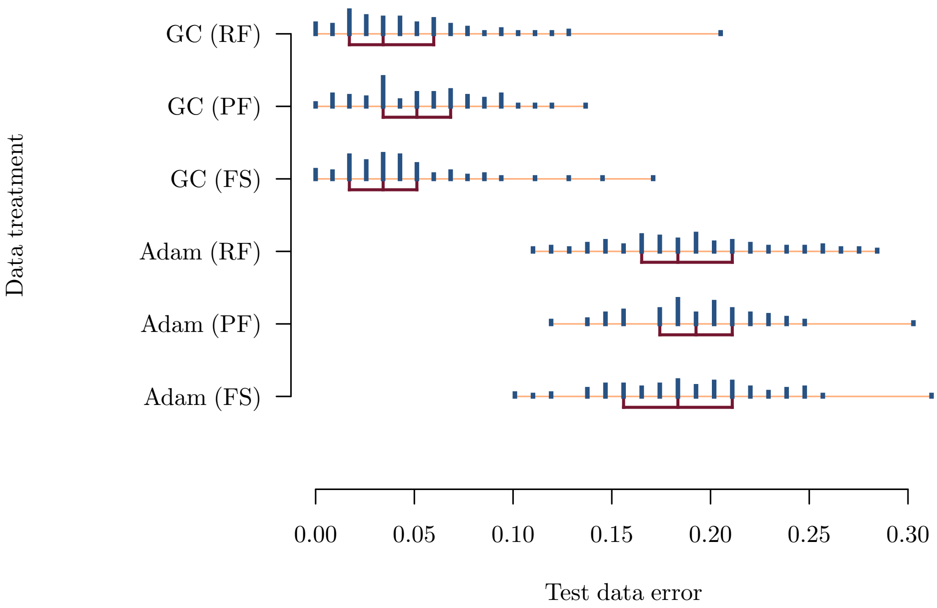
Figure 6.9: Discrimination results for GC mice and Adam et al. (2002) datasets using varying peak quantification methods.

6.4 Comparison of feature selection and classification methods

The optimal model and feature selection methods in discrimination are now considered. This is determined in two stages: to begin, the classification model is kept constant to compare feature selection methods and then the converse, the classification models are compared using the previously best performing feature selection method.

6.4.1 Feature selection

Figure 6.10 displays the effect of feature selection on the classification using SVM. From the figure, there is no obviously preferred feature selection candidate. However, RF and FS had lower 75th, 50th and 25th percentiles of test data error proportions for the GC mice dataset than PF. It might therefore be assumed that the type of feature selection method employed does not alter the outcome greatly for these datasets. From a computational point of view, the FS feature selection method would be preferable since it is much less computationally intensive.



Spectra normalisation	Classification method	Feature selection	Peak quantification	Data transformation	Imputation method
EQN	SVM	⟨variable⟩	Peak area	$\log_2$	Peak minimum

Figure 6.10: Discrimination results for GC mice and Adam et al. (2002) datasets using varying feature selection methods.

Despite there being no obvious 'best' feature selection method on the test data prediction error results, investigation into that relationship, in conjunction with the number of features chosen for the final model on the test data (for all 100 re-samplings), may reveal a preferred feature selection method. Figure 6.11 and Table 6.2 provide information about the relationship between the predictive test error, feature selection method and the number of features selected.

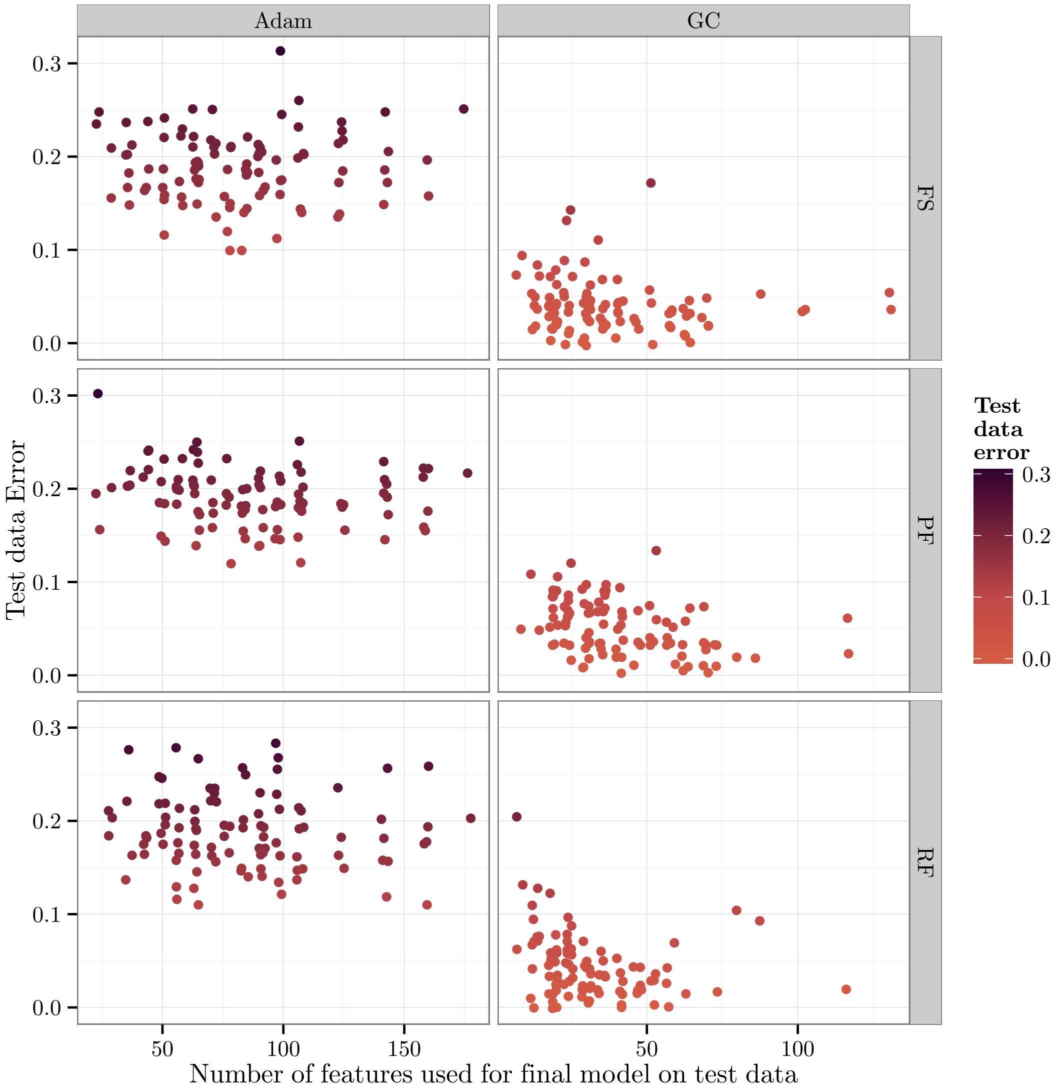


Figure 6.11: Number of features selected for the final predictive model in the 100 iterations. The number of features is paired with the resulting test data error. Slight perturbation of points was used to reveal obscured points.

Table 6.2 shows that PF feature selection generally forces the final predictive model to include more features to inform its classification. As PF does not produce better classification than the other methods, it may be inferred it selects redundant features which do not effectively discriminate the disease groups. On the other hand, RF applied to the GC mice dataset selected the least number of features in the final classification model on the test data (Table 6.2). It can be seen in Figure 6.11 that RF feature selection provided a generally lower predictive error in the 25 to 50 feature-range than both FS and PF. Unfortunately, RF may have been marred by the 'Ockham's razor' decision process in model optimisation: when two or more potential models produce the same error on the training data the model using fewer features is considered 'optimal' (§6.1). This occurred in 82 of the 100 iterations,[5] and is why RF selected fewer than 25 features more often than FS and PF.

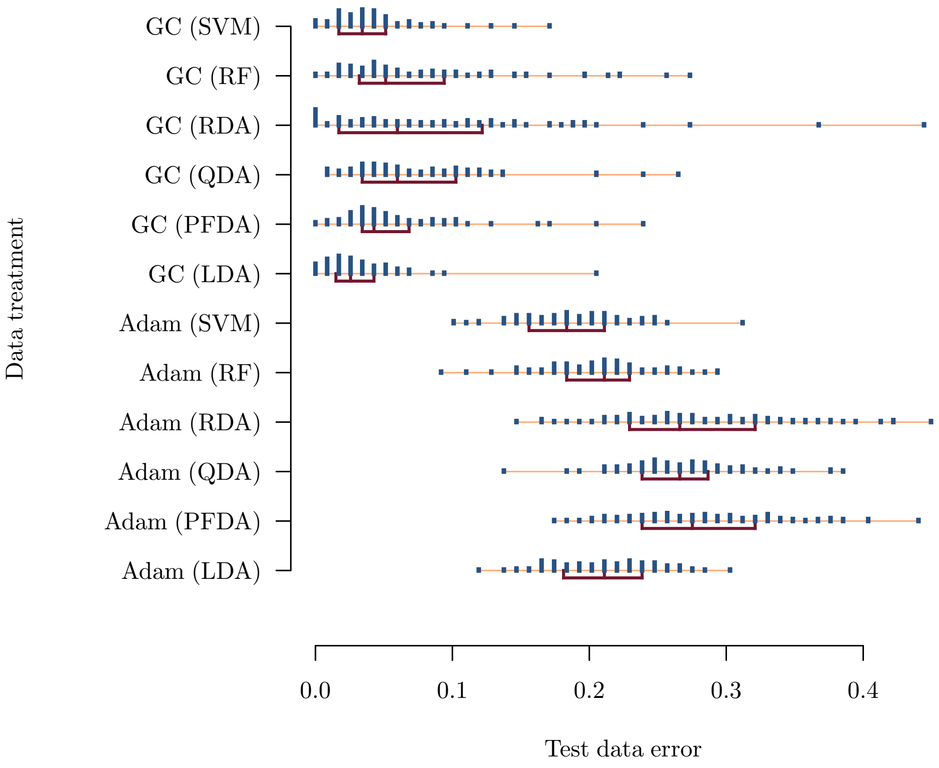
Table 6.2: The interquartile range of the number of features used in the test data prediction model for the Adam et al. (2002) and GC mice datasets.

Feature selection	Adam			GC		
method	Q1	Q2	Q3	Q1	Q2	Q3
FS	57	77	98	19	30	48
PF	64	84	107	24	35	54
RF	57	84	98	19	24	41

6.4.2 Classification model

Using FS to rank features, Figure 6.12 sets out the results from the discrimination methods outlined in §5.3, §5.4 and §5.5. For both the GC mice and Adam et al. (2002) datasets, SVM and LDA provided the lowest prediction error. RF error proportions were similar to those of SVM but had slightly greater variability. As was expected, the statistical methods extended for high dimensional data (RDA and PFDA, §5.4) did not produce error proportions as low as traditional methods. RDA and PFDA are more appropriate for datasets with a greater 'feature to sample' ratio. Both these methods had larger variability in the error proportions, a result of their inherent compromise of pragmatic prediction with biased parameter estimation (Guo et al., 2007) that make these models more susceptible to the choice training data. The flexible assumption of QDA to allow different covariance structures for the classification groups, compared to LDA, showed no benefit. QDA had a median error proportion of 0.060 for the GC mice data, which is more than twice as high than the median for LDA (0.026).

[5] Of these 82 scenarios where two or more models had equal five-fold cross-validation error, 88% of the time (72 of 82) the competing models had the minimum possible five-fold cross-validation error of 0.



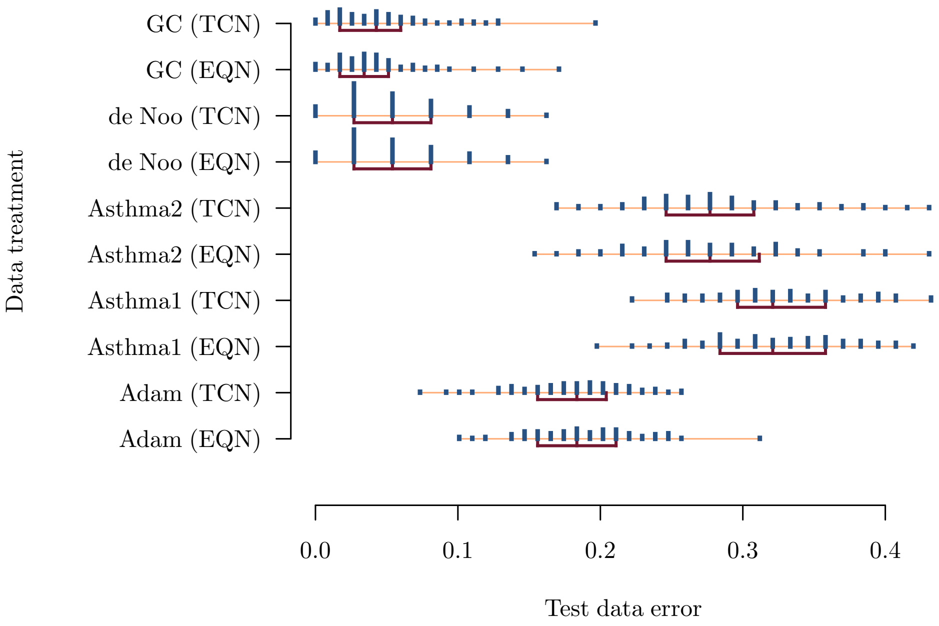
Spectra normalisation	Classification method	Feature selection	Peak quantification	Data transformation	Imputation method
EQN	⟨variable⟩	FS	Peak area	$\log_2$	Peak minimum

Figure 6.12: Discrimination results for GC mice and Adam et al. (2002) datasets using different classification methods.

6.5 The effects of spectra normalisation on classification

As discussed in Chapter 2, there is no recommended best practice in spectra preprocessing. The objective evaluation of methods for spectra normalisation may be performed by observing the classification results when varying the normalisation method, while keeping all other factors constant.

Figure 6.13 depicts the classification results for EQN and TCN applied to all five datasets. Unfortunately, there is no clear advantage for any one method according to the classification results. EQN is slightly better for the GC dataset, with a lower 25th percentile of test data prediction error than TCN. For the Adam et al. (2002) dataset, TCN produced nominally better results with a lower 25th percentile of test data prediction error than EQN, although the 75th percentile and median were equal for both methods. As expected, the normalisation method had little to no effect on classification for the de Noo et al. (2006) dataset (§4.2.2 and §6.3). This is because the data were already pre-processed, as discussed in §1.3.4 and §3.1.4, and have a strong classification signal using a limited number of features. For the asthma1 and asthma2 datasets, the normalisation method made little difference to the classification outcome. The 75th percentile of the test data prediction error was slightly lower when EQN was used in the asthma2 dataset than TCN but this is insufficient evidence to choose one normalisation method over another.



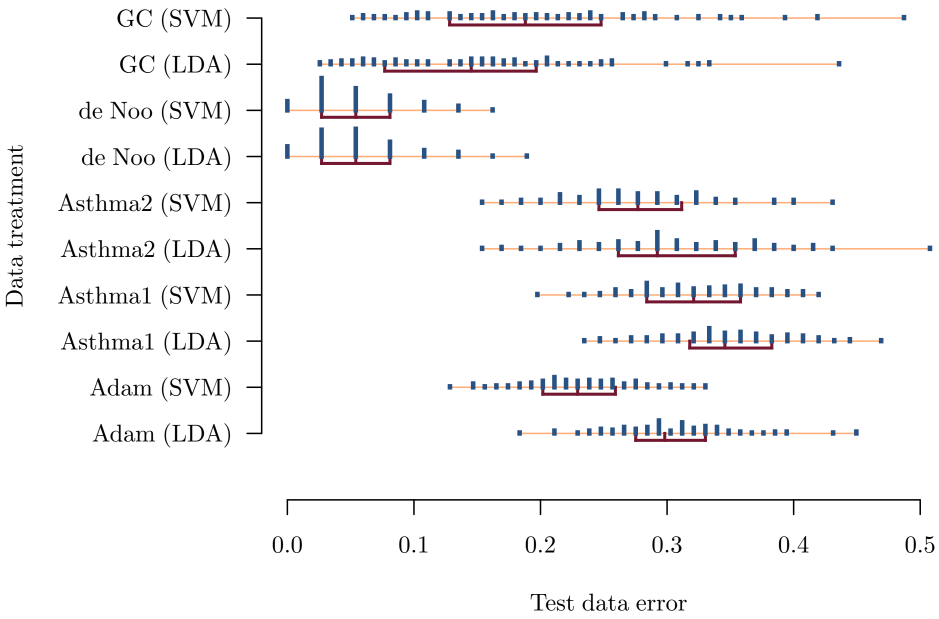
Spectra normalisation	Classification method	Feature selection	Peak quantification	Data transformation	Imputation method
⟨variable⟩	SVM	FS	Peak area	$\log_2$	Peak minimum

Figure 6.13: Discrimination results for GC mice and Adam et al. (2002) datasets using different spectra normalisation methods.

As there is no preferred normalisation method for discrimination, the final results in the next section will use EQN which has good theoretical control over peak variability and is likely to be a good method to handle highly variable spectra in proteomic datasets generally. TCN would be equally useful for these datasets according to the prediction results on the test datasets.

6.6 Comparison of proteomic data results

Now that an 'optimal' process for expression construction and analysis elements has been chosen, a final assessment of discriminating methods for the datasets can be conducted. In this section, the full granulated outcome variable was used rather than the dichotomised diseased and control groups. The GC dataset therefore has five diagnostic groups and the Adam et al. (2002) dataset has four diagnostic groups. This increases the complexity of the discrimination models (especially for SVM) and requires more precise discrimination rules to correctly classify cancer and control subgroups. As a result, the prediction error is likely to increase, as is evident in Figure 6.14. FS feature selection was used in conjunction with LDA and SVM classification.



Spectra normalisation	Classification method	Feature selection	Peak quantification	Data transformation	Imputation method
EQN	SVM or LDA	FS	Peak area	$\log_2$	Peak minimum

Figure 6.14: The final discrimination results using SVM and LDA classification.

For the de Noo et al. (2006) data, the classification results were effectively the same for both LDA and SVM classifiers. This can be attributed to the differential peak expressions at numerous m/z locations with very little class-overlap. As discussed previously, this may not be due to true biological signal. SVM classification can be seen to perform better than LDA on both the asthma datasets but the discrimination error is not far from chance (asthma1 and asthma2 datasets having female to male ratios of 117:126 and 96:99, respectively). The results for the asthma datasets are expected as the spectra have not shown any promising signal in biomarker identification analyses.

Discrimination on the Adam et al. (2002) data performed best using SVM. The predictive error did not dramatically increase when the number of disease groups to classify was increased from two to four. The median predictive error moved from less than 0.200 in previous analyses to 0.229 for the SVM classification. The results are better than chance but well below the results of the original Adam et al. (2002) paper. Using the more robust and sophisticated discrimination in this thesis, the Adam et al. (2002) results are not reproducible (Diamandis, 2003; Semmes et al., 2005; McLerran et al., 2008a,b).

Part of the success of LDA in the GC dataset may be attributed to the number of replicates available. With a dataset containing 360 peak expression vectors with 160 features available to be used, the higher-dimensional transformation of the feature space using a non-linear kernel SVM is unnecessary and at the same time, the parameter estimates in LDA draw on a large number of replicates. The median predictive error was below 0.200 for both the SVM and LDA models, much better than chance. However the best performing discrimination combination of FS and LDA with a median predictive error of 0.145 had a less impressive 25th percentile of 0.197, with a maximum of 0.436. This suggests the classification is still sensitive to the training data in model optimisation. The classification using the dichotomy of GC and control classes showed promising low predictive error that lends itself to the potential of diagnostic testing, but the ability to discriminate between the five GC and control subgroups is not satisfactory.

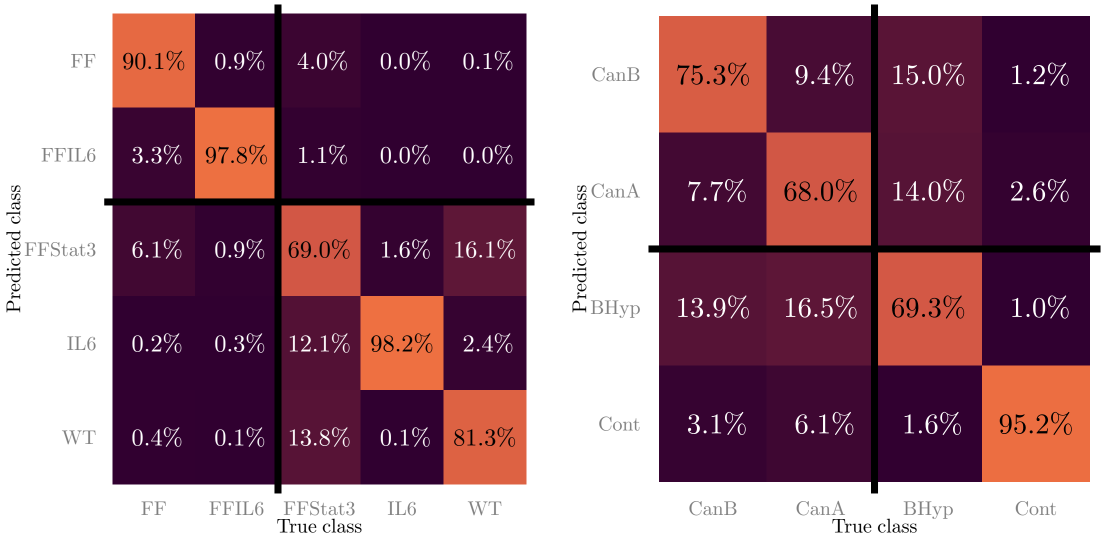
The ability of the discrimination models to correctly classify cancer and control subgroups on the test data is shown in Figure 6.15. These are the confusion matrices (also called contingency matrices; Fawcett, 2006) for classification on the test data using LDA for the GC mice data and SVM for the Adam et al. (2002) data. Each cell in these confusion matrices represents the percentage of peak expression vectors that were classified to certain disease status by the discrimination model (‘predicted class’ on the y-axis). The ‘true class’ is on the x-axis of these plots, therefore each column in each panel represents the distribution of predicted classes for peak expression vectors conditional on underlying disease status.

Ideally, each diagonal entry in Figure 6.15 would be 100%, indicating perfect classification. However, this is not the case. For example the FFStat3 (control) mice in the GC mice data were misclassified 31.0%. Interestingly, FFStat3 mice were correctly predicted to be FFStat3 mice or the other two control mice subtypes, IL6 or WT, 94.9% of the time. In fact, the LDA classification model correctly predicted GC subtypes as GC mice, or control subtypes as control mice, 93.4% or more of the time despite the increase in discrimination categories to five classes.

Table 6.3: Proportional representation of peaks in the 100 final classification models used to classify the GC mice test data, restricted to peaks occurring in 80 or more models.

Peak	Potential	Proportion of models
m/z	biomarker[†]	peak is in classifier[††]
3959		1.00
7412	**	1.00
7490		1.00
7806		1.00
7917		1.00
8302		1.00
8533	**	1.00
8607	**	1.00
8717		1.00
9305	**	1.00
13648	**	1.00
14836		1.00
14993		1.00
16654		1.00
17458	**	1.00
12161	**	0.99
6821	**	0.98
6076		0.96
4650		0.95
5204	**	0.95
8337	**	0.94
8418		0.94
9214		0.94
9239		0.94
8831	**	0.93
8970		0.93
8067		0.91
15844		0.91
4993		0.88
7738		0.87
9319		0.87
8505		0.86
9712		0.85
6602	**	0.84
5453		0.83
5752		0.83
6354		0.82
9059		0.80

[†] Potential biomarker is defined as peaks identified in the linear models of §4.2.1 that are shown in Figure 4.10 and Table C.1. [††] Length of maroon bar represents proportion of models in which the peak is selected. The opacity of the colour represents the proportion of observed values within the peak prior to imputation (high opacity indicates low missingness).



(a) GC mice data using LDA

(b) Adam et al. data using SVM

Figure 6.15: Confusion matrices as heatmaps of the average test predictive error for the GC and Adam et al. (2002) datasets. Columns represent the empirical conditional probabilities of model predicted classes given the underlying disease classification.

The SVM discrimination models for the Adam et al. (2002) data were less successful. Control patients were correctly classified 95.2% of the time and was the group most often correctly predicted. The benign hyperplasia (BHyp) group was misclassified most often at 30.7% of the time. The model incorrectly classified BHyp patients as cancer (CanA or CanB) patients 29.1% of the time, and similarly, CanA and CanB patients were incorrectly classified as BHyp patients 13.9% and 16.5% of the time, respectively. From the confusion matrix for the Adam et al. (2002) data for discriminating the four classes, the strongest classification signal is between the control patients and the benign hyperplasia or more progressed prostate abnormality patients. In fact, the sensitivity and specificity was 96.4% and 95.2% respectively, for a test between those dichotomised diagnoses.

To complete the biomarker identification for the GC mice dataset (§4.2), the peaks used in the final discrimination models are given in Table 6.3. The important peaks used in the LDA models to classify the disease status of the test data were consistent with 38 peaks used in 80 or more of the 100 final models. Of these 38 peaks, 12 were previously identified potential biomarkers in Figures 4.10 and 4.11 in §4.2. The potential biomarkers that were not deemed to be good discriminators between classification groups (not used often in the final models) were recognisably those peaks that had a high proportion of missingness, for example peaks at 8867, 14421 and 15631m/z. The median number of peaks used in the final models was 62; the top 62 'most included peaks' across all the final discrimination models accounted for 82.2% of the total features used. The observed stability in the features selected is important and indicates the classification models were not sensitive to the training data allocation, at the same time providing a valuable ranked set of potential biomarkers.

6.7 Summary

The process undertaken in this chapter has been a conditional step-wise approach to optimise the factors that affect discrimination. The approach taken is not guaranteed to find the optimal set of expression construction and analysis elements, and could be explored in future work. However, the final set of factors selected produced promising results, especially in the GC mice dataset, which obtained a test data prediction error proportion of less than 0.150 using only eight mice in each of the five groups.

Table 6.4 summarises the steps and decisions to 'optimise' the set of expression construction and analysis elements. There were clear advantages using certain methods over others. Peak area almost uniformly produced a lower discrimination error on test data for the GC mice and Adam et al. (2002) datasets compared to the standard method of maximum peak intensity (peak height). Methods for imputing missing values are an important area for future research. Missing values are a large stumbling block for most analysis methods that generally do not natively handle missing values. Peak minimum was deemed the best imputation method, although there is plenty of scope for this to be improved. It is argued here that the minimum method is a sensible combination of simplicity, theoretical reasoning (missingness generally occurs through sub-detection threshold expression) and relatively improved discrimination.

Two factors that were observed to have little impact on the discrimination results were the feature selection method and the normalisation method. These observations are considered provisional and would benefit from investigation on other datasets. Furthermore, the dimension reduction method PCA actually hindered the discrimination between disease groups using the classification models on the test data studied here.

The choice of model should be data dependent. LDA performed better within the more traditional paradigm where the data contain more observations than features (GC mice data) and SVM performed better in the SnLp type setting (i.e., for the remaining datasets). SVMs require additional parameter optimisation over traditional statistical techniques but are often worth the extra computational work, as observed on all datasets except the GC mice data.

The GC mice data performed exceptionally well in classification. There were notable improvements in discrimination if only one additional replicate per subject was involved in model building. In addition to the beneficial effect of increasing replication, the GC mice data demonstrated that pooling or averaging spectra can improve and stabilise results. It is speculated this is induced by a combination of reducing the variability that comes with proteomic data and the use of replicates to cover missing values. Although the GC mice data results need to be interpreted in the context of being in a murine model, the results are promising and potentially diagnostic.

Table 6.4: The step-wise approach of Chapter 6 to 'optimise' the discrimination of classification groups.

	Expression construction and analysis elements					
Analysis step	Spectra normalisation	Classification method	Feature selection	Peak quantification	Data transformation	Imputation method
Starting parameters	EQN	SVM	PF	Peak height	$\log_2$	Median
Expression averaging[†]	↓	↓	↓	↓	↓	↓
No. of replicates[†]	↓	↓	↓	↓	↓	↓
Data transformation	↓	↓	↓	↓	⟨variable⟩	↓
Imputation	↓	↓	↓	↓	$\log_2$	⟨variable⟩
Peak quantification	↓	↓	↓	⟨variable⟩	↓	Min/missMDA
Feature selection	↓	↓	⟨variable⟩	Peak area	↓	Min
Discrim. method	↓	⟨variable⟩	FS	↓	↓	↓
Spectra normalisation	⟨variable⟩	SVM/LDA	↓	↓	↓	↓
Final parameters	EQN	SVM/LDA	FS	Peak area	$\log_2$	Min

[†] GC mice dataset specific analysis.

Chapter 7

Concluding remarks

Bioinformatics is a tremendously challenging field that requires collaboration from many disciplines to optimise experimental success. With the continued development of MALDI TOF-MS technology and mathematical methods in tandem, there is much scope for improved sensitivity in biomarker detection, and the potential for developing diagnostic methods. The potential biomarkers for gastric cancer identified here require further biological investigation.

Statisticians can provide important guidance with respect to proteomic MS experimental design, especially as people with an intimate knowledge of the model requirements and assumptions encountered at the analysis stage. Although SVA and RUV are available to estimate additional sources of unexplained variability in peak expression data, statistical input at the design stage would ensure measurable variables of experimental subjects are recorded and explicitly modelled in the resulting proteomic MS data analysis. Information on non-experimental variables, such as patient demographics, in the human proteomic MS datasets were in large part not available here. The inclusion of all variables that are potential confounders at the analysis stage would result in more robust analyses, greater understanding of the biological processes and increased generalisability.

Two observations from this work are worth emphasising.

- Replicate spectra derived from the same experimental subjects are extremely beneficial and should be undertaken whenever feasible. Because of the systematic variability due to batch effects, replication is required to obtain precise disease effect estimates. The classification in Chapter 6 demonstrated that the inclusion of a single replicate led to considerably improved discrimination.

- A more sophisticated way to impute missing values is required in this context. Supervised methods for biomarker identification would be appropriate, and

> unsupervised methods would be required in diagnostic settings. Both scenarios would certainly be bolstered by the availability of additional variables relating to experimental subjects.

The reproducibility of the MALDI TOF-MS data is one of the largest hurdles in the development of a diagnostic test and indeed reliable biomarker identification. The results from well designed experiments performed in varied locations will provide indications of the true capability of the technology and current mathematical methods.

Appendix A

R code

All files referenced in the current appendix are available in the `R/` directory at:

`https://github.com/tystan/thesis/`.

A.1 Morphological operators

Description	File	Functions
Naive erosion and top-hat:		
	/00_erosion_slow.R	erode(), dilate(), tophat()
Line segment erosion:		
	/01_erosion_quick.R	erode_quick()
Naive erosion for unequally spaced values:		
	/02_cts_erosion_slow.R	erode_cts_slow()
Continuous line segment erosion:		
	/03_cts_erosion_quick.R	erode_cts_quick()

A.2 Spectra normalisation

Description	File	Functions
Empirical quantile normalisation:		
	/04_quant_norm.R	quant_norm()
Pairwise spectra MA normalisation:		
	/05_ma_adj.R	ma_adj()

A.3 Peak alignment

Description	File	Functions
Calculate W matrix for an N- and M-alignment:		
	/06_create_w.R	w_matrix()
Dendrogram peak alignment:		
	/07_dendro_peak_align.R	dendro_peak_align()

A.4 Surrogate variable analysis

Description	File	Functions
Get SVA adjusted expression matrix:		
	/08_do_sva.R	do_sva()

A.5 Pairwise fusion linear discriminant analysis

Description	File	Functions
Create a PFDA object:		
	/09_create_pfda_obj.R	create_pfda_obj()
Predict class for new data and a PFDA object:		
	/10_pfda_predict.R	pfda_predict()

A.6 Pareto Fronts for variable ranking

Description	File	Functions
Calculate dominating features:		
	/11_dom_feat.c	dom_feat()
Pareto Front wrapper functions:		
	/12_pareto_fronts.R	pareto_ranking()

Appendix B

Continuous line segment algorithm proof

Proof of Proposition 2.1. This proof considers the three possible ways $r_{\text{CLSA}}\left(f\left(x_i\right)\right)$ is calculated, separately. From the CLSA definition, the three cases can be seen below:

$$r_{\text{CLSA}}\left(f\left(x_i\right)\right)=\begin{cases} g\left(x_{w_i^\vartriangle}\right) & \text{if } \theta_{w_i^\triangledown}=\theta_{w_i^\vartriangle+1} \\ h\left(x_{w_i^\triangledown}\right) & \text{if } \theta_{w_i^\triangledown-1}=\theta_{w_i^\vartriangle} \\ \min\left\{g\left(x_{w_i^\vartriangle}\right),h\left(x_{w_i^\triangledown}\right)\right\} & \text{otherwise.} \end{cases}$$

For ease of reference, the value θ_i, will be referred to as 'the block number for x_i'.

Please note that because $0\leq x_{w_i^\vartriangle}-x_{w_i^\triangledown}\leq k$ (the difference of the extreme values contained in the moving window for point x_i), therefore $0\leq\theta_{w_i^\vartriangle}-\theta_{w_i^\triangledown}\leq 1$. This implies $\theta_{w_i^\vartriangle}-\theta_{w_i^\triangledown}\in\{0,1\}$ as the vector Θ contains non-decreasing integer values.

Case 1: $\theta_{w_i^\triangledown}=\theta_{w_i^\vartriangle+1}$.

Case 1 implies the block number of the lower bound index for x_i has the same block number as the upper bound index for x_i. Therefore $\theta_{w_i^\triangledown}=\theta_i=\theta_{w_i^\vartriangle}$, denote this value as θ_1^*.

If the block numbers of $x_{w_i^\vartriangle}$ and $x_{w_i^\vartriangle+1}$ are equal, this means that the coverage of the moving window to the left includes the first instance of θ_1^* in Θ, i.e.,

$$\theta_{w_i^\triangledown-1}<\theta_{w_i^\triangledown}=\theta_1^*, \tag{$*$}$$

because the moving window for x_i must include either the first instance of θ_1^* in the coverage to the left side of the moving window or the last instance of θ_1^* in the coverage to the right side of the moving window. As $\theta_{w_i^\triangledown}=\theta_{w_i^\vartriangle+1}$, this prohibits the latter from being true. Therefore,

$$\begin{aligned} r_{\text{CLSA}}\left(f\left(x_i\right)\right) &= g(x_{w_i^\vartriangle}) \\ &= \min\left[g(x_{w_i^\vartriangle-1}),f(x_{w_i^\vartriangle})\right] \\ &= \min\left[\min\left[g(x_{w_i^\vartriangle-2}),f(x_{w_i^\vartriangle-1})\right],f(x_{w_i^\vartriangle})\right] \\ &= \min\left[g(x_{w_i^\vartriangle-2}),f(x_{w_i^\vartriangle-1}),f(x_{w_i^\vartriangle})\right] \\ &\qquad\text{because of the idempotence of minimum} \\ &= \min\left[g(x_{w_i^\triangledown}),f(x_{w_i^\triangledown+1}),\ldots,f(x_{w_i^\vartriangle-1}),f(x_{w_i^\vartriangle})\right] \\ &= \min\left[f(x_{w_i^\triangledown}),f(x_{w_i^\triangledown+1}),\ldots,f(x_{w_i^\vartriangle-1}),f(x_{w_i^\vartriangle})\right] \\ &\qquad\text{because } g(x_{w_i^\triangledown})=f(x_{w_i^\triangledown}) \text{ for } \theta_{w_i^\triangledown-1}<\theta_{w_i^\triangledown}\ldots(*) \\ &= \left\{\min f(x_t):x_i-k_0\leq x_t\leq x_i+k_0\right\} \\ &= \epsilon_B(f)(x_i). \end{aligned}$$

Case 2: $\theta_{w_i^{\triangledown}-1} = \theta_{w_i^{\triangle}}$.

Case 2 implies the block number of the lower bound index for x_i has the same block number as the upper bound index for x_i. Therefore $\theta_{w_i^{\triangledown}} = \theta_i = \theta_{w_i^{\triangle}}$, denote this value as θ_2^*.

If the block numbers of $x_{w_i^{\triangledown}-1}$ and $x_{w_i^{\triangledown}}$ are equal, this means that the coverage of the moving window to the right includes the last instance of θ_2^* in Θ, i.e.,

$$\theta_2^* = \theta_{w_i^{\triangle}} < \theta_{w_i^{\triangle}+1}, \tag{$**$}$$

because the moving window for x_i must include either the first instance of θ_2^* in the coverage to the left side of the moving window or the last instance of θ_2^* in the coverage to the right side of the moving window. As $\theta_{w_i^{\triangledown}-1} = \theta_{w_i^{\triangle}}$, this prohibits the former from being true. Therefore,

$$\begin{aligned}
r_{\text{CLSA}}\left(f\left(x_i\right)\right) &= h(x_{w_i^{\triangledown}}) \\
&= \min\left[f(x_{w_i^{\triangledown}}), h(x_{w_i^{\triangledown}+1})\right] \\
&= \min\left[f(x_{w_i^{\triangledown}}), \min\left[f(x_{w_i^{\triangledown}+1}), h(x_{w_i^{\triangledown}+2})\right]\right] \\
&= \min\left[f(x_{w_i^{\triangledown}}), f(x_{w_i^{\triangledown}+1}), h(x_{w_i^{\triangledown}+2})\right] \\
&\qquad\qquad \text{because of the idempotence of minimum} \\
&= \min\left[f(x_{w_i^{\triangledown}}), f(x_{w_i^{\triangledown}+1}), \ldots, f(x_{w_i^{\triangle}-1}), h(x_{w_i^{\triangle}})\right] \\
&= \min\left[f(x_{w_i^{\triangledown}}), f(x_{w_i^{\triangledown}+1}), \ldots, f(x_{w_i^{\triangle}-1}), f(x_{w_i^{\triangle}})\right] \\
&\qquad\qquad \text{because } h(x_{w_i^{\triangle}}) = f(x_{w_i^{\triangle}}) \text{ for } \theta_{w_i^{\triangle}} < \theta_{w_i^{\triangle}+1} \ldots(**) \\
&= \left\{\min f(x_t) : x_i - k_0 \leq x_t \leq x_i + k_0\right\} \\
&= \epsilon_B(f)(x_i).
\end{aligned}$$

Case 3: $\theta_{w_i^{\triangledown}-1} \neq \theta_{w_i^{\triangle}}$ and $\theta_{w_i^{\triangledown}} \neq \theta_{w_i^{\triangle}+1}$.

Case 3 implies the block numbers of the upper and lower bounds of x_i are different, i.e. $\theta_{w_i^{\triangle}} \neq \theta_{w_i^{\triangledown}}$. Therefore there is an integer b_i such that $w_i^{\triangledown} \leq b_i < b_i + 1 \leq w_i^{\triangle}$ and $\theta_{w_i^{\triangledown}} = \theta_{b_i} < \theta_{b_i+1} = \theta_{w_i^{\triangle}}$. The CLSA minimum for case 3 is then calculated using,

$$r_{\text{CLSA}}\left(f\left(x_i\right)\right) \;=\; \min\left[h(x_{w_i^{\triangledown}}), g(x_{w_i^{\triangle}})\right],$$

where

$$\begin{aligned}
h(x_{w_i^{\triangledown}}) &= \min\left[f(x_{w_i^{\triangledown}}), f(x_{w_i^{\triangledown}+1}), \ldots, f(x_{b_i-1}), f(x_{b_i})\right] \\
&\qquad\qquad \text{because } h(x_{b_i}) = f(x_{b_i}) \text{ for } \theta_{b_i} < \theta_{b_i+1}\text{, and} \\
g(x_{w_i^{\triangle}}) &= \min\left[f(x_{b_i+1}), f(x_{b_i+2}), \ldots, f(x_{w_i^{\triangle}-1}), f(x_{w_i^{\triangle}})\right] \\
&\qquad\qquad \text{because } g(x_{b_i+1}) = f(x_{b_i+1}) \text{ for } \theta_{b_i} < \theta_{b_i+1}.
\end{aligned}$$

$$\begin{aligned}
\therefore r_{\text{CLSA}}\left(f\left(x_i\right)\right) &= \min\left[f(x_{w_i^{\triangledown}}), f(x_{w_i^{\triangledown}+1}), \ldots, f(x_{w_i^{\vartriangle}-1}), f(x_{w_i^{\vartriangle}})\right] \\
&= \{\min f(x_t) : x_i - k_0 \leq x_t \leq x_i + k_0\} \\
&= \epsilon_B(f)(x_i).
\end{aligned}$$

As $\theta_{w_i^{\vartriangle}} - \theta_{w_i^{\triangledown}} \in \{0, 1\}$, all cases have been considered and it has been shown,

$$r_{\text{CLSA}}\left(f\left(x_i\right)\right) = \epsilon_B\left(f\right)\left(x_i\right) \quad \forall x_i \in X.$$

□

Appendix C

Biomarker investigation supplementary information

C.1 PCA plots for GC mice dataset

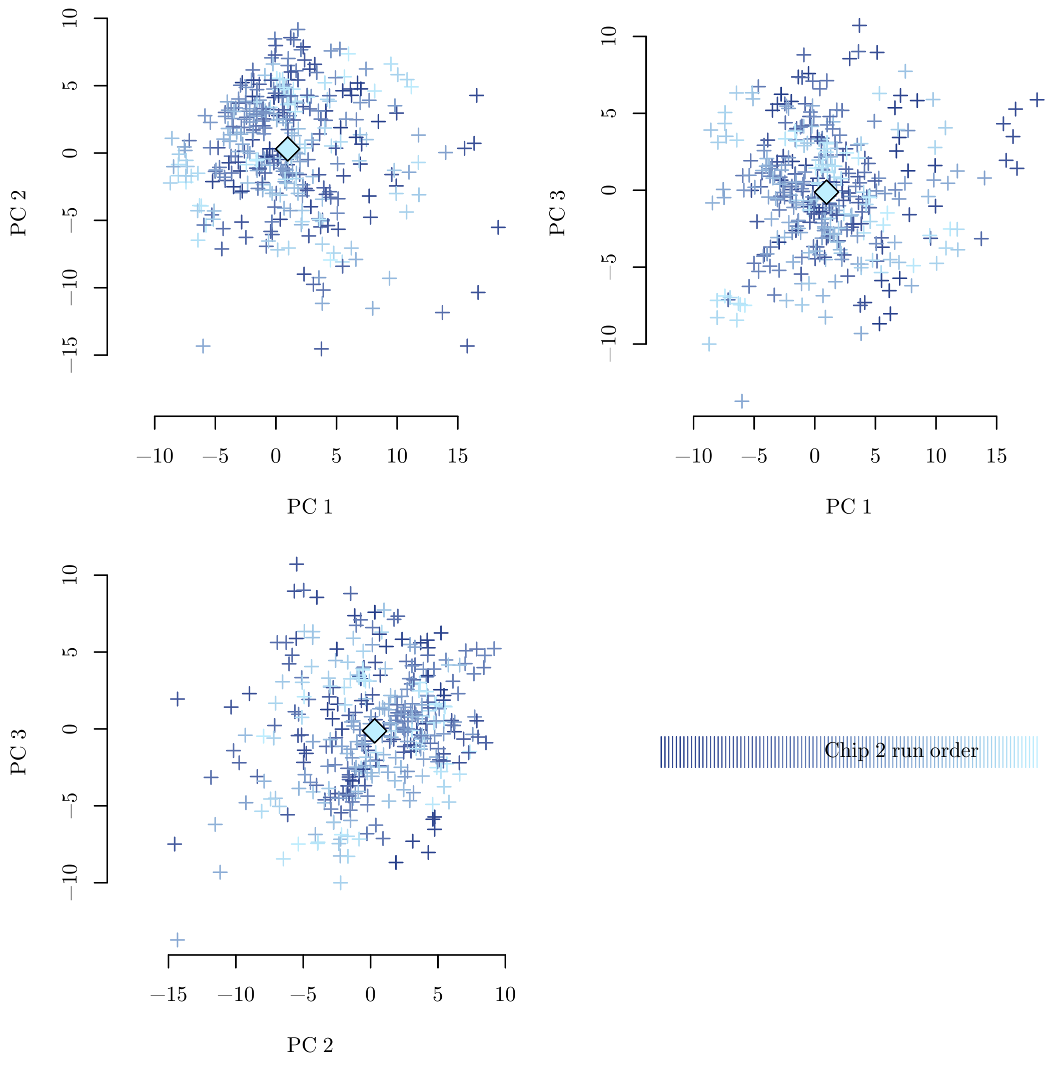


Figure C.1: PCA plots of GC mice peak expression data by chip 2 run-order. The 1080 PCA points are plotted in a random order irrespective of chip run-order to avoid a visual bias from plotting points in run-order.

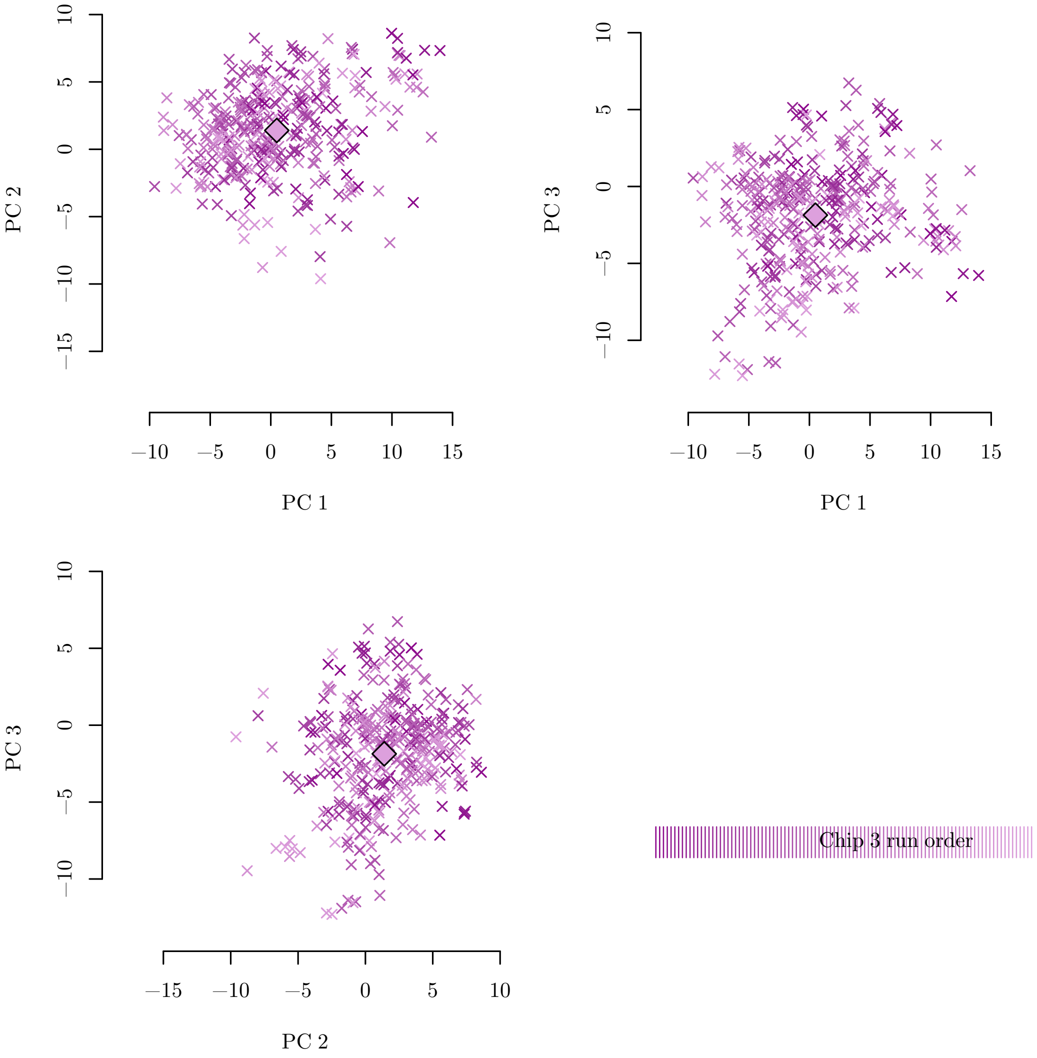


Figure C.2: PCA plots of GC mice peak expression data by chip 3 run-order. The 1080 PCA points are plotted in a random order irrespective of chip run-order to avoid a visual bias from plotting points in run-order.

C.2 PCA plots for asthma datasets

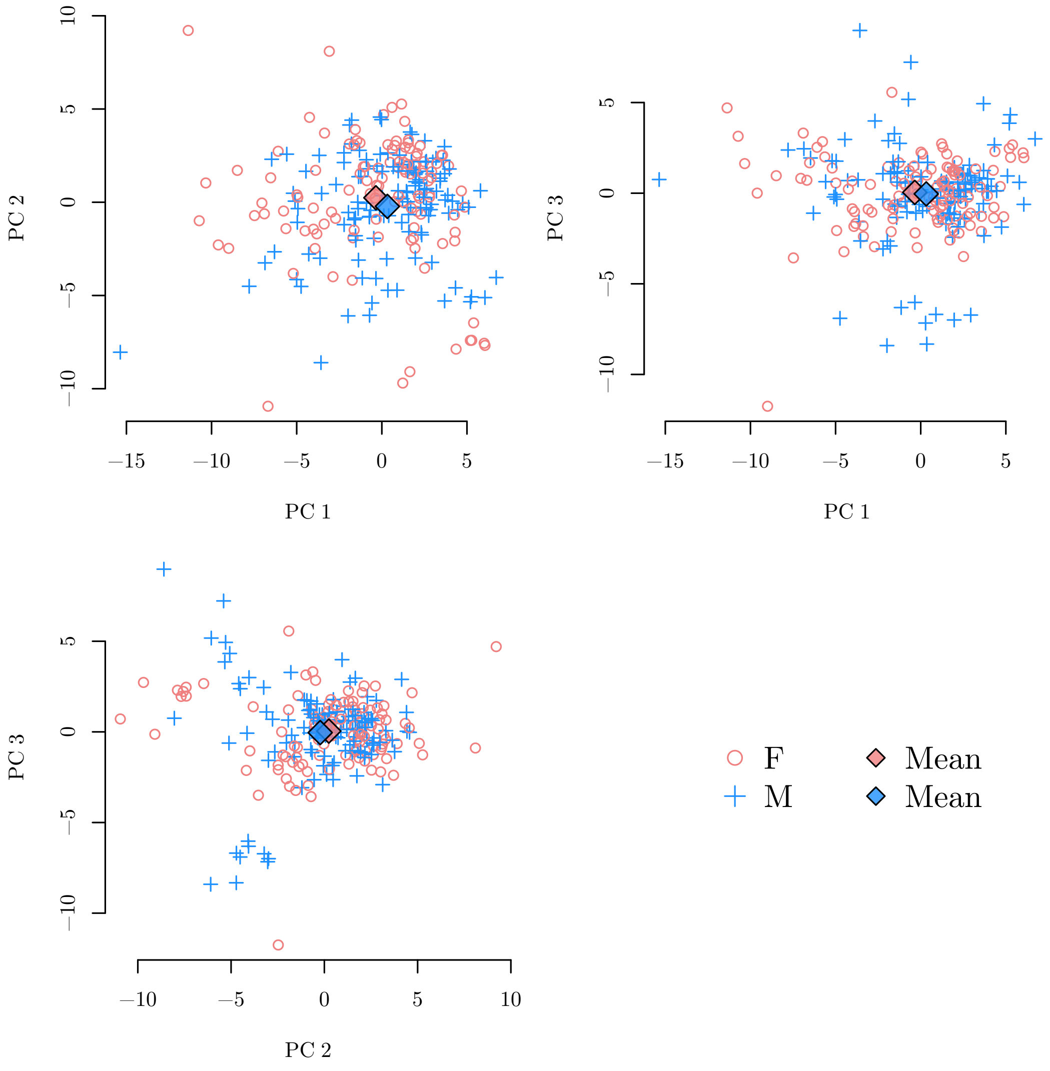


Figure C.3: PCA plots of peak expression intensities of the asthma1 dataset by group labels. The 243 PCA points are plotted in a random order irrespective of group membership to avoid a visual bias from plotting points in group order.

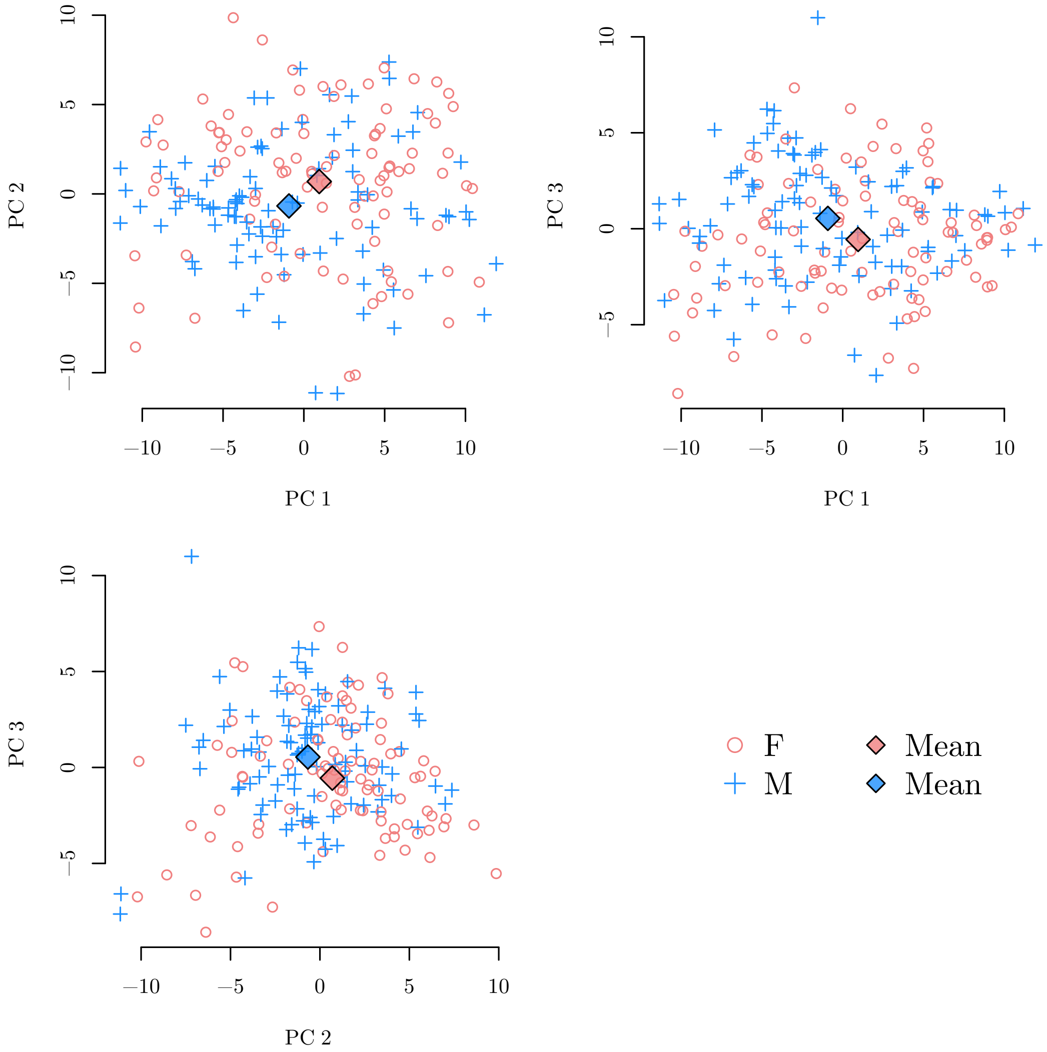
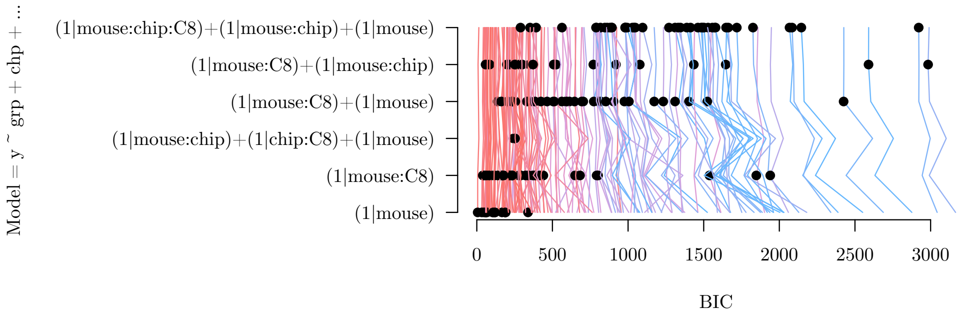
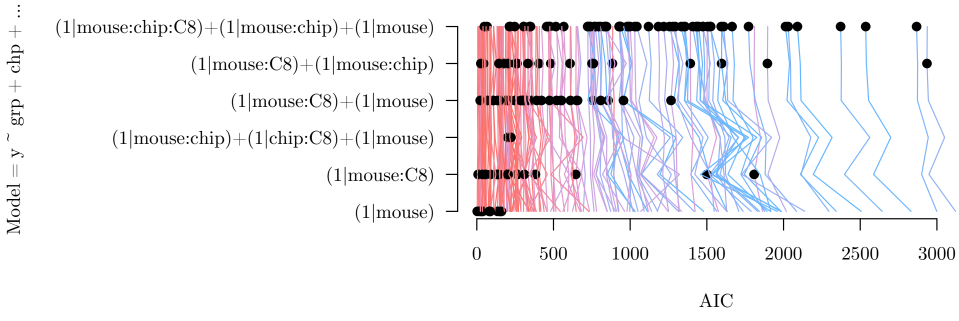


Figure C.4: PCA plots of peak expression intensities of the asthma2 dataset by group labels. The 197 PCA points are plotted in a random order irrespective of group membership to avoid a visual bias from plotting points in group order.

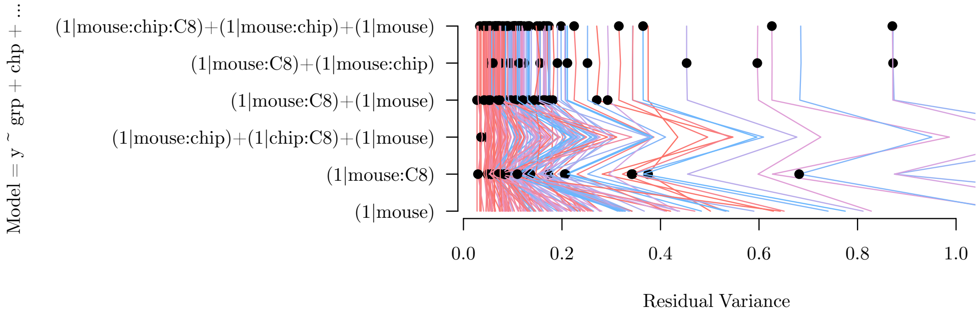
C.3 Supplementary biomarker investigation for the GC mice dataset



(a) BIC



(b) AIC



(c) Residual variance

Figure C.5: Statistics to assess different random effects model structures. Each line represents the (a) BIC, (b) AIC and (c) residual variance for a particular peak ($p = 1, 2, \ldots, 159$) over each of the six different models. The colour of the line indicates the missingness in the peak and the black dot represents the minimum statistic for that peak over the six models, i.e. the algorithmically preferred model.

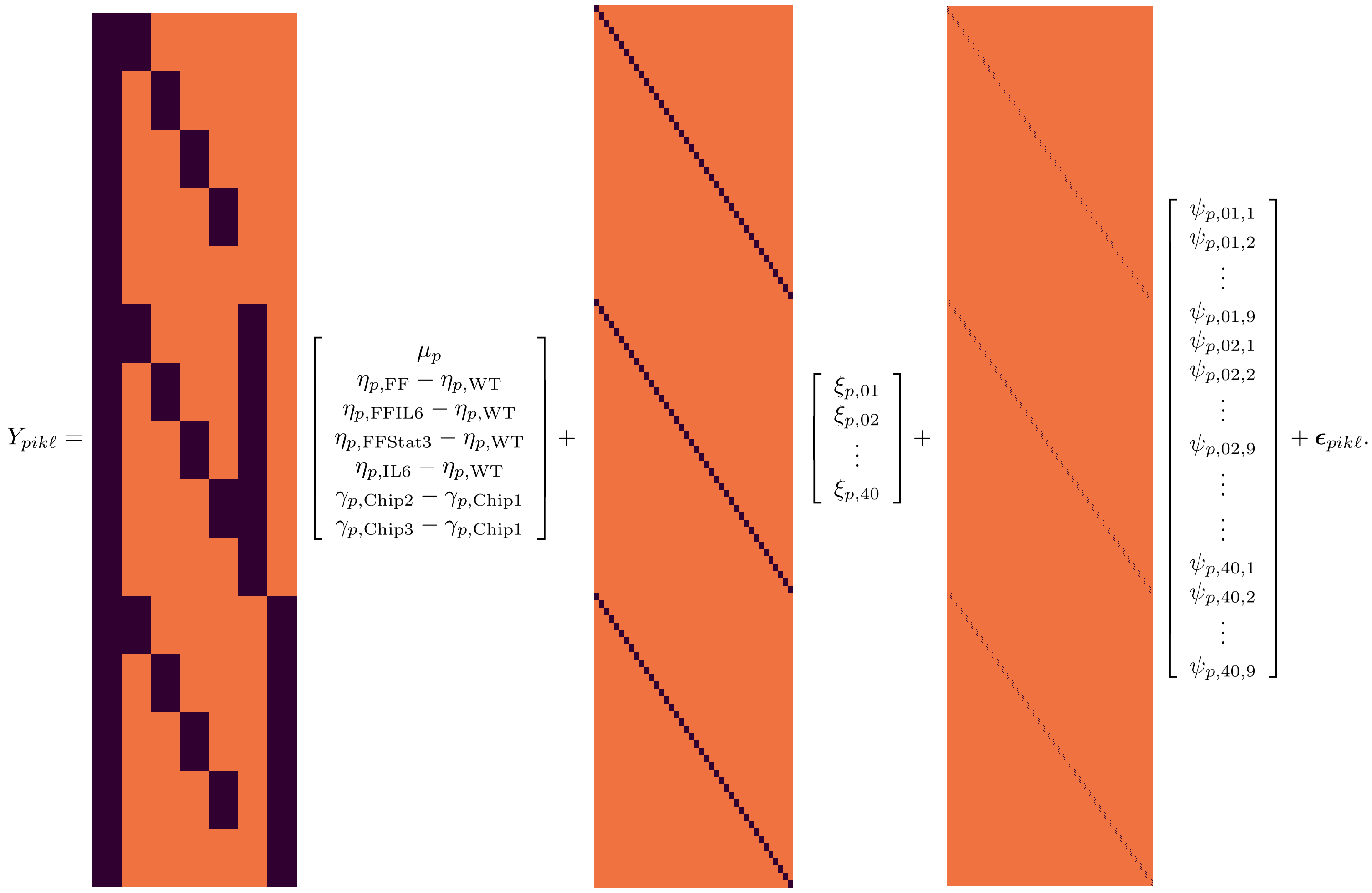

$Y_{pik\ell} =$
μ_p
$\eta_{p,\text{FF}} - \eta_{p,\text{WT}}$
$\eta_{p,\text{FFIL6}} - \eta_{p,\text{WT}}$
$\eta_{p,\text{FFStat3}} - \eta_{p,\text{WT}}$
$\eta_{p,\text{IL6}} - \eta_{p,\text{WT}}$
$\gamma_{p,\text{Chip2}} - \gamma_{p,\text{Chip1}}$
$\gamma_{p,\text{Chip3}} - \gamma_{p,\text{Chip1}}$
$+$
$\xi_{p,01}$
$\xi_{p,02}$
$\vdots$
$\xi_{p,40}$
$+$
$\psi_{p,01,1}$
$\psi_{p,01,2}$
$\vdots$
$\psi_{p,01,9}$
$\psi_{p,02,1}$
$\psi_{p,02,2}$
$\vdots$
$\psi_{p,02,9}$
$\vdots$
$\vdots$
$\psi_{p,40,1}$
$\psi_{p,40,2}$
$\vdots$
$\psi_{p,40,9}$
$+ \boldsymbol{\epsilon}_{pik\ell}.$

Figure C.6: Schematic of the three-level LME model structure for the GC mice data of the form $\mathbf{Y}_p = \mathbf{X}\boldsymbol{\beta}_p + \mathbf{Z}\mathbf{B}_p + \boldsymbol{\epsilon}_p$ where $\mathbf{Z}$ has been partitioned into $\mathbf{Z}_2$ and $\mathbf{Z}_1$ corresponding to the third and second level random effects, respectively. Orange cells in the design matrices denote 0 values, dark maroon 1 values.

Table C.1: Peaks with significant group difference and fold changes of at least $\frac{3}{2}$ in the GC mice peak expression dataset. Fold change is relative to the control group: up regulation implies higher expression in the cancer group. The p-values are multiple comparison adjusted using the Benjamini & Hochberg method to maintain a FDR at 0.05.

m/z	Fold change	Up/down regulated	p-value
6602	1.55	+	2.6×10^{-3}
6821	1.92	+	5.7×10^{-4}
7412	1.69	-	5.4×10^{-8}
7806	1.71	-	2.6×10^{-4}
8337	1.92	+	7.9×10^{-6}
8533	1.65	-	4.4×10^{-6}
8607	1.79	-	9.9×10^{-6}
8831	1.69	+	9.8×10^{-5}
8867	1.55	-	7.4×10^{-3}
9305	1.88	+	1.2×10^{-4}
12161	1.84	+	1.6×10^{-4}
13648	2.60	+	4.3×10^{-6}
14421	1.69	+	4.4×10^{-5}
14836	1.54	-	2.4×10^{-8}
17458	2.16	-	4.4×10^{-6}

Table C.2: Missingness observed for peak expressions for peaks 8337, 8607 and 12161m/z for GC mice dataset.

Peak		Chip[†]			
(m/z)	Group	1	2	3	Total
8337	FF	36	37	36	109
	FFIL6	0	0	0	0
	FFStat3	53	53	54	160
	IL6	0	0	0	0
	WT	72	69	71	212
	Total	161	159	161	481
8607	FF	33	40	42	115
	FFIL6	7	11	21	39
	FFStat3	7	8	10	25
	IL6	0	0	0	0
	WT	16	14	9	39
	Total	63	73	82	218
12161	FF	5	2	0	7
	FFIL6	25	0	0	25
	FFStat3	19	0	0	19
	IL6	60	27	24	111
	WT	16	15	14	45
	Total	125	44	38	207

[†] Each chip and group cell has the maximum availability of 72 peak expressions from the 72 spectra taken from each group on each chip.

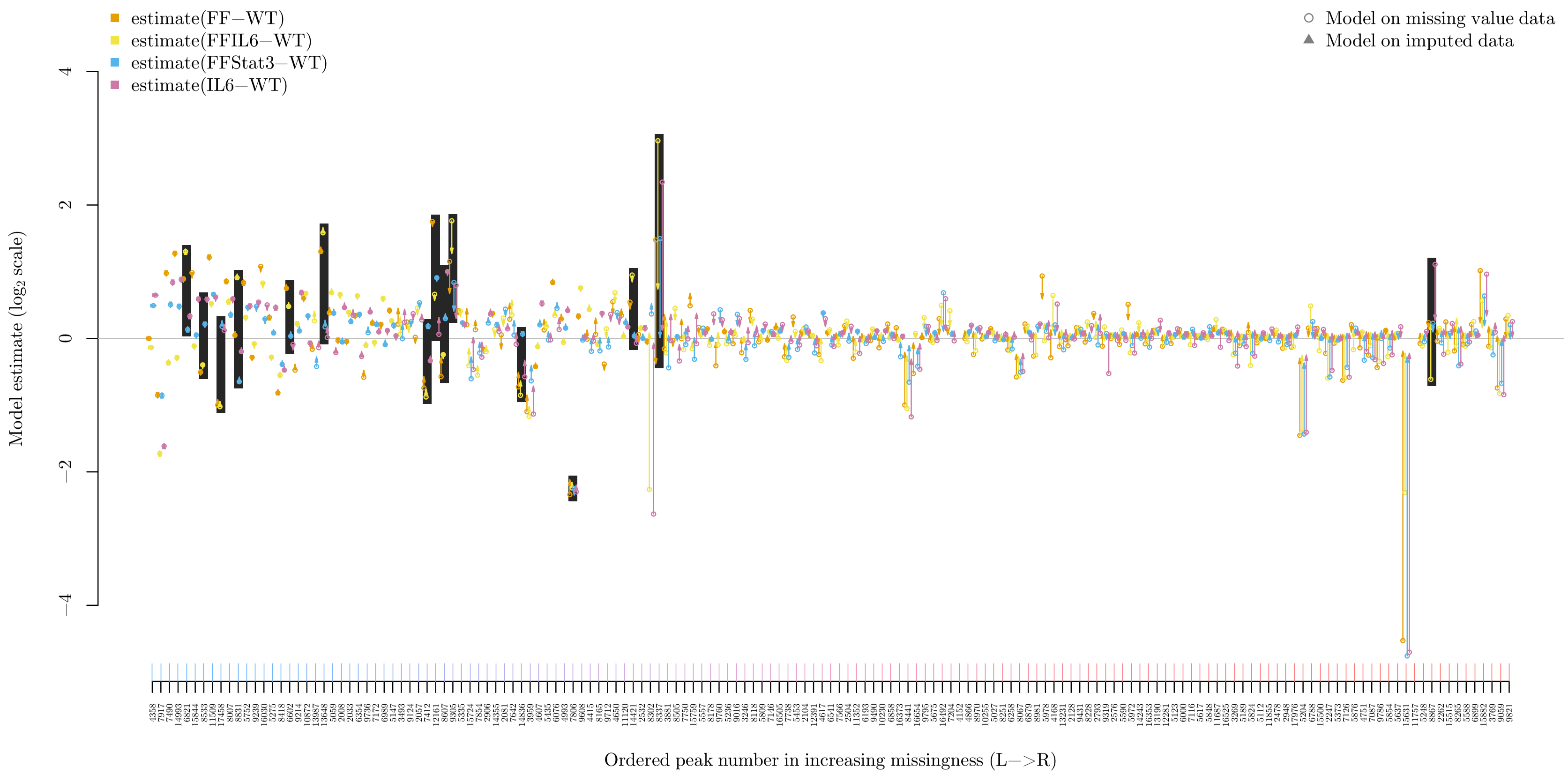


Figure C.7: Parameter estimates, relative to the WT group, for the LME models when missing values are ignored and when the missing values are kNN imputed.

C.4 Supplementary biomarker investigation for the de Noo et al. (2006) and Adam et al. (2002) datasets

Table C.3: Peaks with significant group difference and fold changes of at least 2 in the de Noo et al. (2006) peak expression dataset. Fold change is relative to the control group: up regulation implies higher expression in the cancer group. The p-values are multiple comparison adjusted using the Benjamini & Hochberg method to maintain a FDR at 0.05.

		Up/down	
m/z	Fold change	regulated	p-value
1210	2.22	-	2.7×10^{-11}
1266	2.40	-	1.3×10^{-14}
1337	3.22	-	2.9×10^{-11}
1352	2.79	-	$< 10^{-16}$
1417	2.66	-	3.8×10^{-3}
1437	5.05	-	2.0×10^{-5}
1691	2.47	+	6.7×10^{-13}
1781	3.63	+	$< 10^{-16}$
1800	4.64	+	4.1×10^{-4}
1849	2.36	+	3.2×10^{-12}
1868	6.03	+	$< 10^{-16}$
1886	4.43	+	3.5×10^{-6}
1897	2.42	+	1.1×10^{-16}
2019	2.48	+	3.5×10^{-9}
2024	3.12	+	1.2×10^{-6}
3193	2.01	-	4.5×10^{-11}
3267	2.04	-	6.3×10^{-13}
4056	2.16	+	$< 10^{-16}$

Table C.4: Peaks with significant group difference and fold changes of at least $\frac{3}{2}$ in the Adam et al. (2002) peak expression dataset. Fold change is relative to the control group: up regulation implies higher expression in the cancer group. The p-values are multiple comparison adjusted using the Benjamini & Hochberg method to maintain a FDR at 0.05.

		Up/down	
m/z	Fold change	regulated	p-value
2145	1.65	-	4.2×10^{-4}
2502	1.54	+	2.7×10^{-2}
3281	1.98	-	1.2×10^{-7}
3964	2.06	-	1.2×10^{-7}
4070	1.78	+	5.8×10^{-6}
4250	1.84	+	1.7×10^{-5}
4291	1.98	-	1.3×10^{-7}
4499	1.84	-	1.8×10^{-9}
4580	1.56	+	2.7×10^{-3}
4603	1.76	-	1.8×10^{-2}
4690	1.60	+	7.7×10^{-3}
5997	2.74	+	6.0×10^{-3}
7442	1.68	+	8.8×10^{-5}
7687	1.57	+	9.5×10^{-3}
8142	1.58	+	$< 10^{-16}$
8293	1.89	+	1.4×10^{-7}
8354	1.71	+	$< 10^{-16}$
9150	1.64	+	3.2×10^{-5}

C.5 Supplementary biomarker investigation for the asthma datasets

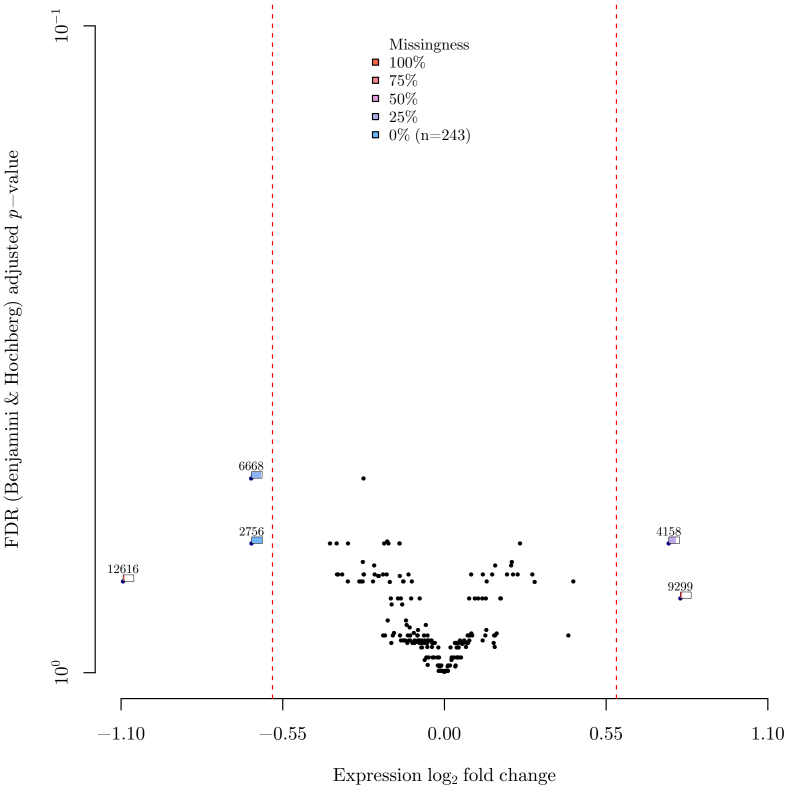


Figure C.8: Volcano plots for asthma1 dataset relating to group differences peak expressions; adjusted p-value vs. fold change on the $\log_2$-scale. Missingness observed for each peak is indicated by rectangle fill adjacent to point.

Table C.5: Peaks with fold changes of at least $\frac{3}{2}$ in the asthma1 peak expression dataset. Fold change is relative to the female group: up regulation implies higher expression in the male group. The p-values are multiple comparison adjusted using the Benjamini & Hochberg method to maintain a FDR at 0.05.

		Up/down	
m/z	Fold change	regulated	p-value
2756	1.58	-	0.631
4158	1.70	+	0.631
6668	1.58	-	0.500
9299	1.74	+	0.768
12616	2.13	-	0.722

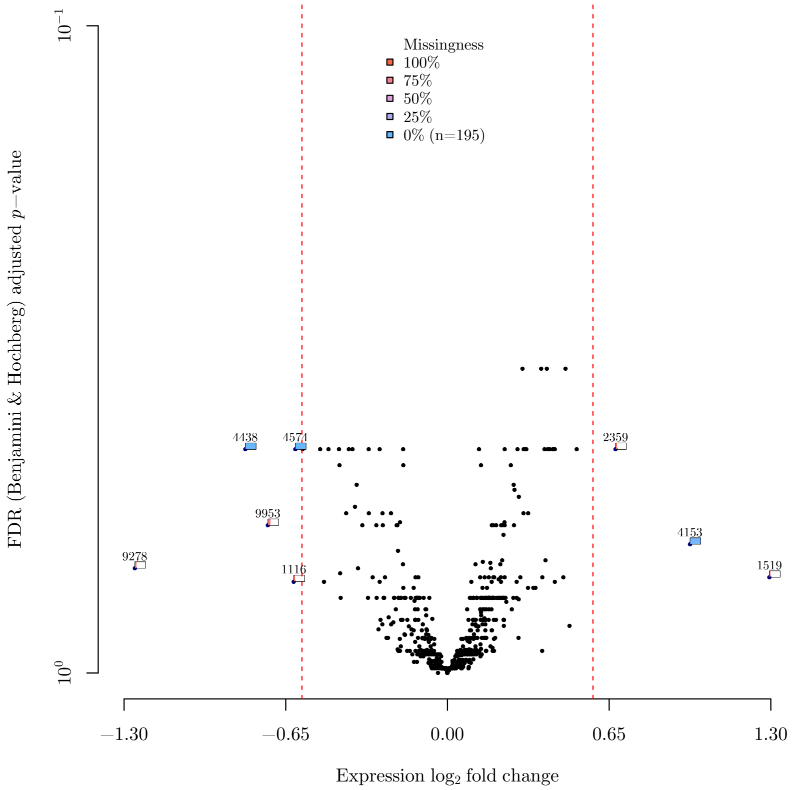


Figure C.9: Volcano plots for asthma2 dataset relating to group differences peak expressions; adjusted p-value vs. fold change on the $\log_2$-scale. Missingness observed for each peak is indicated by rectangle fill adjacent to point.

Table C.6: Peaks with fold changes of at least $\frac{3}{2}$ in the asthma2 peak expression dataset. Fold change is relative to the female group: up regulation implies higher expression in the male group. The p-values are multiple comparison adjusted using the Benjamini & Hochberg method to maintain a FDR at 0.05.

		Up/down	
m/z	Fold change	regulated	p-value
1116	1.53	-	0.722
1519	2.45	+	0.711
2359	1.60	+	0.451
4153	1.97	+	0.632
4438	1.76	-	0.451
4574	1.53	-	0.451
9278	2.39	-	0.689
9953	1.65	-	0.591

Appendix D

Probabilistic LDA classification of GC mice

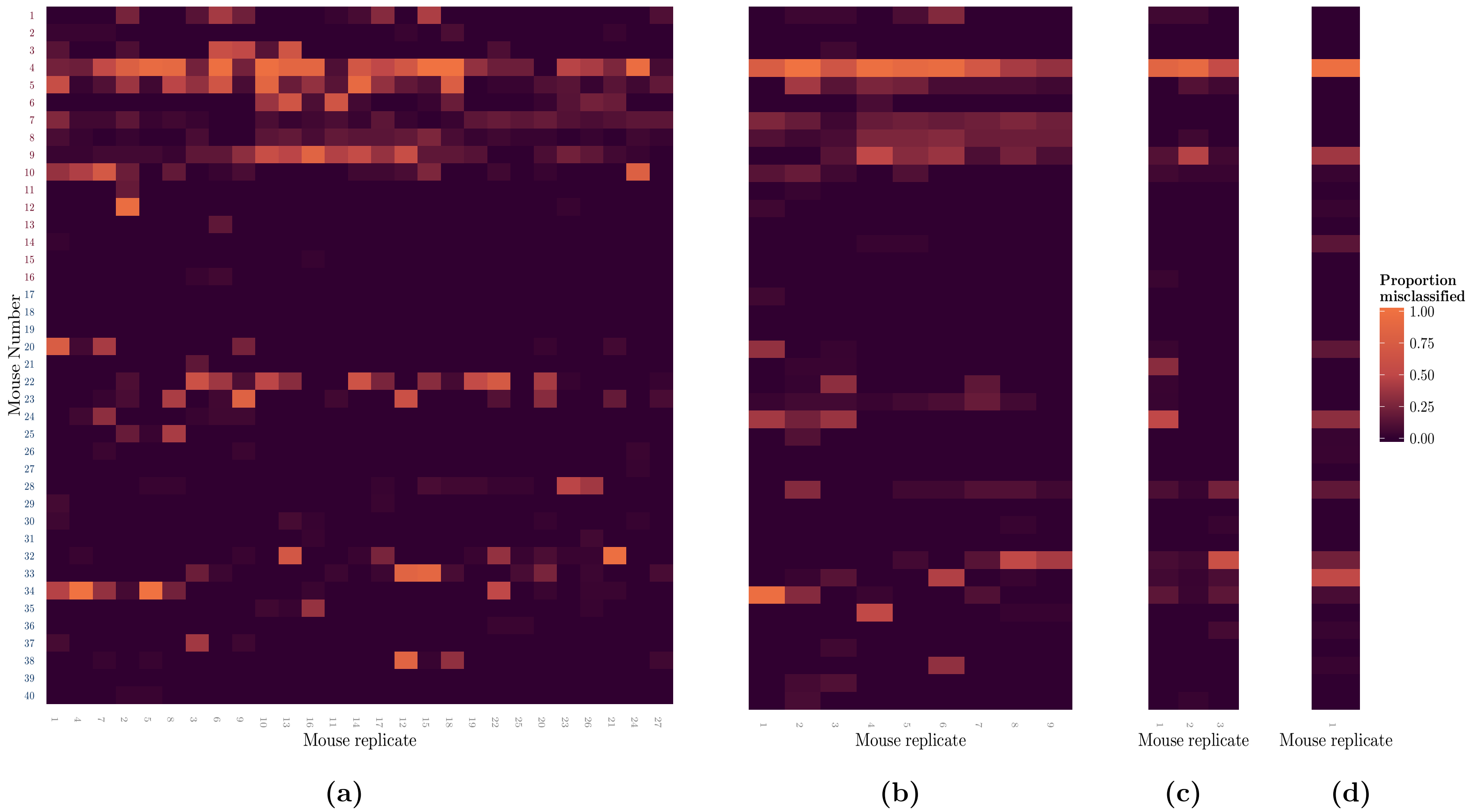


Figure D.1: Heatmap of the LDA predictive error for GC mice dataset for different peak expression vector averaging schemes: (a) none, (b) C8, (c) aliquot and (d) mouse.

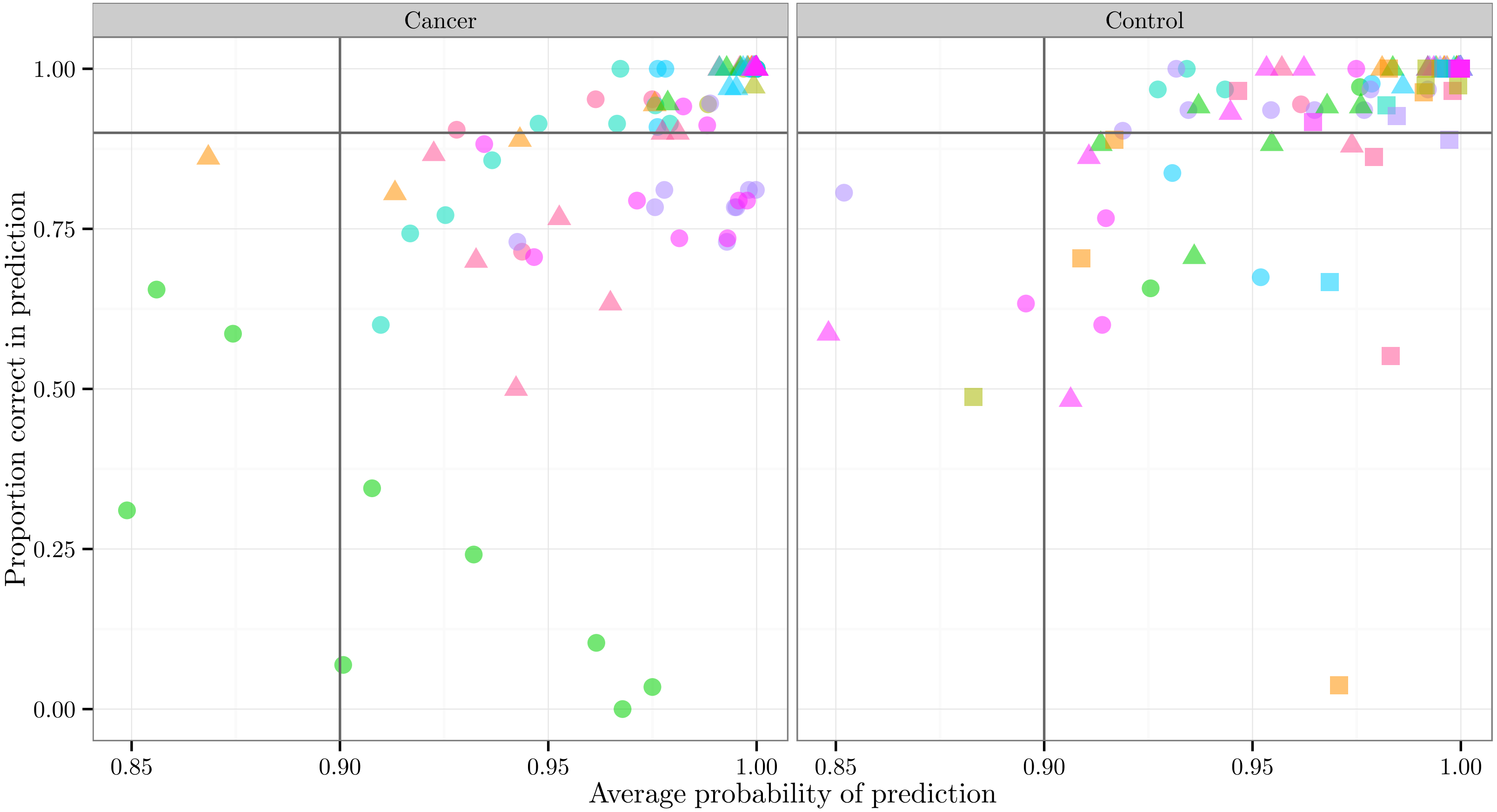


Figure D.2: Relationship between prediction certainty and prediction outcome for the C8 averaged data using LDA classification. Each distinct colour/character represents the nine replicates of the mice within the GC and control groups. The vertical line represents the average certainty/probability of 0.90 for that particular mouse/replicate and the horizontal line represents 90% correct prediction.

Bibliography

AABB. Circular of information for the use of human blood and blood components. Technical Report 133011, AABB (formerly American Association of Blood Banks), the American Red Cross, America's Blood Centers, and the Armed Services Blood Program, Maryland USA, 2013. URL http://www.aabb.org/tm/coi/Documents/coi1113.pdf.

B.-L. Adam, Y. Qu, J. W. Davis, M. D. Ward, O. J. Semmes, P. F. Schellhammer, Y. Yasui, Z. Feng, G. L. Wright Jr., M. A. Clements, and L. H. Cazares. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research*, 62(13):3609–3614, 2002.

AIHW. Health system expenditure on cancer and other neoplasms in australia 2008-09. Technical Report Cancer series 81. Cat. no. CAN 78, Australian Institute of Health and Welfare, Canberra: AIHW, 2013. URL http://www.aihw.gov.au/cancer-publications/.

J. Albrethsen. Reproducibility in protein profiling by MALDI-TOF mass spectrometry. *Clinical Chemistry*, 53(5):852–858, 2007.

J. Albrethsen. The first decade of MALDI protein profiling: A lesson in translational biomarker research. *Journal of Proteomics*, 74(6):765–773, 2011.

T. Alexandrov, J. Decker, B. Mertens, A. M. Deelder, R. A. E. M. Tollenaar, P. Maass, and H. Thiele. Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation. *Bioinformatics*, 25(5):643–649, 2009.

D. B. Allison, X. Cui, G. P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7(1): 55–65, 2006.

M. Anderle, S. Roy, H. Lin, C. Becker, and K. Joho. Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum. *Bioinformatics*, 20(18):3575–3582, 2004.

N. L. Anderson and N. G. Anderson. The human plasma proteome: History, character, and diagnostic prospects. *Molecular & Cellular Proteomics*, 1(11):845–867, 2002.

R. Armananzas, Y. Saeys, I. Inza, M. Garcia-Torres, C. Bielza, Y. Van de Peer, and P. Larranaga. Peakbin selection in mass spectrometry data using a consensus approach with estimation of distribution algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(3):760–774, 2011.

V. Audigier, F. Husson, and J. Josse. A principal components method to impute missing values for mixed data. *ArXiv e-prints*, Jan 2013.

K. A. Baggerly, J. S. Morris, and K. R. Coombes. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics*, 20(5):777–785, 2004.

K. A. Baggerly, J. S. Morris, S. R. Edmonson, and K. R. Coombes. Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. *Journal of the National Cancer Institute*, 97(4):307–309, 2005.

K. V. Ballman, D. E. Grill, A. L. Oberg, and T. M. Therneau. Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics*, 20(16):2778–2786, 2004. doi: 10.1093/bioinformatics/bth327.

N. Barbarini and P. Magni. Accurate peak list extraction from proteomic mass spectra for identification and profiling studies. *BMC bioinformatics*, 11(1):518, 2010.

D. Bates. *Computational methods for mixed models*, 2011. `http://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf`.

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *ArXiv e-print (arXiv:1406.5823)*, 2014a. Submitted to the *Journal of Statistical Software.*

D. Bates, M. Maechler, B. Bolker, and S. Walker. *lme4: Linear mixed-effects models using Eigen and S4*, 2014b. R package version 1.1-7.

D. M. Bates. *lme4: Mixed-effects modeling with R.* Prepublication, 2010. URL `http://http://lme4.r-forge.r-project.org/book/`.

W. C. Bauldry. *Introduction to Real Analysis, in Introduction to Real Analysis: An Educational Approach.* John Wiley & Sons, Inc, 2009.

R. C. Beavis and B. T. Chait. High-accuracy molecular mass determination of proteins using matrix-assisted laser desorption mass spectrometry. *Analytical Chemistry*, 62(17):1836–1840, 1990.

R. C. Beavis and B. T. Chait. Matrix-assisted laser desorption ionization mass-spectrometry of proteins. *Methods in enzymology*, 270:519–551, 1996.

R. C. Beavis, B. T. Chait, and H. M. Fales. Cinnamic acid derivatives as matrices for ultraviolet laser desorption mass spectrometry of proteins. *Rapid Communications in Mass Spectrometry*, 3(12):432–435, 1989a.

R. C. Beavis, B. T. Chait, and K. G. Standing. Factors affecting the ultraviolet laser desorption of proteins. *Rapid Communications in Mass Spectrometry*, 3(7): 233–237, 1989b.

R. C. Beavis, B. T. Chait, and K. G. Standing. Matrix-assisted laser-desorption mass spectrometry using 355 nm radiation. *Rapid Communications in Mass Spectrometry*, 3(12):436–439, 1989c.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

S. L. Berger, T. Kouzarides, R. Shiekhattar, and A. Shilatifard. An operational definition of epigenetics. *Genes & Development*, 23(7):781–783, 2009.

Bio-Rad Laboratories, Accessed July 2010. `http://www.bio-rad.com/`.

B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

S. Bolton and C. Bon. *Pharmaceutical statistics: practical and clinical applications*, chapter Statistical analysis of biomarkers from -omics technologies, pages 437–442. CRC Press, 2009.

H. W. Borchers. *pracma: Practical Numerical Math Functions*, 2012. URL `http://CRAN.R-project.org/package=pracma`. R package version 1.3.3.

S. Borman, H. Russell, and G. Siuzdak. A mass spec timeline. *Today's Chemist at Work*, 2003.

L. Breiman. Technical note: Some properties of splitting criteria. *Machine Learning*, 24(1):41–47, 1996.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

L. Breiman. Setting up, using, and understanding random forests V4.0. Technical report, University of California, Berkeley, 2002. URL `http://oz.berkeley.edu/users/breiman/Using_random_forests_v4.0.pdf`.

M. U. A. Bromba and H. Ziegler. Application hints for Savitzky-Golay digital smoothing filters. *Analytical Chemistry*, 53(11):1583–1586, 1981.

A. Buja and N. Eyuboglu. Remarks on parallel analysis. *Multivariate behavioral research*, 27(4):509–540, 1992.

A. K. Callesen, R. Christensen, J. S. Madsen, W. Vach, E. Zapico, S. Cold, P. E. Jørgensen, O. Mogensen, T. A. Kruse, and O. N. Jensen. Reproducibility of serum protein profiling by systematic assessment using solid-phase extraction and matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Communications in Mass Spectrometry*, 22(3):291–300, 2008.

N. A. Campbell, J. B. Reece, and N. Meyers. *Biology*. Pearson Education Australia, 2006.

C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:1–27, 2011.

P.-H. Chen, R.-E. Fan, and C.-J. Lin. A study on SMO-type decomposition methods for support vector machines. *IEEE Transactions on Neural Networks*, 17:893–908, 2006.

Y. Chen, M. Mori, A. C. Pastusek, K. A. Schug, and P. K. Dasgupta. On-line electrodialytic salt removal in electrospray ionization mass spectrometry of proteins. *Analytical Chemistry*, 83(3):1015–1021, 2011.

V. Cherkassky and F. M. Mulier. *Learning from data: concepts, theory, and methods*. Wiley, 2007.

T. P. Conrads, V. A. Fusaro, S. Ross, D. Johann, V. Rajapakse, B. A. Hitt, S. M. Steinberg, E. C. Kohn, D. A. Fishman, G. Whitely, J. C. Barrett, L. A. Liotta, E. F. Petricoin, and T. D. Veenstra. High-resolution serum proteomic features for ovarian cancer detection. *Endocrine Related Cancer*, 11(2):163–178, 2004.

K. R. Coombes, K. A. Baggerly, and J. S. Morris. Pre-processing mass spectrometry data. In W. Dubitzky, M. Granzow, and D. Berrar, editors, *Fundamentals of Data Mining in Genomics and Proteomics*, pages 79–102. Springer, 2004.

K. R. Coombes, S. Tsavachidis, J. S. Morris, K. A. Baggerly, M.-C. Hung, and H. M. Kuerer. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, 5(16):4107–4117, 2005.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

X. Cui and G. A. Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*, 4(4):210, 2003.

A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application.* Cambridge University Press, 1997.

M. E. de Noo, B. J. A. Mertens, A. Özalp, M. R. Bladergroen, M. P. J. van der Werff, C. J. H. van de Velde, A. M. Deelder, and R. A. E. M. Tollenaar. Detection of colorectal cancer using MALDI-TOF serum protein profiling. *European journal of cancer*, 42(8):1068–1076, 2006.

M. C. P. de Souto, P. A. Jaskowiak, and I. G. Costa. Impact of missing data imputation methods on gene expression clustering and classification. *BMC Bioinformatics*, 16(1):64, 2015.

K. H. Desai and J. D. Storey. Cross-dimensional inference of dependent high-dimensional data. *Journal of the American Statistical Association*, 107(497):135–151, 2012.

E. P. Diamandis. Proteomic patterns in biological fluids: Do they represent the future of cancer diagnostics? *Clinical Chemistry*, 49(8):1272–1275, 2003.

T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel. *e1071: Misc Functions of the Department of Statistics (e1071)*, 2011. URL `http://CRAN.R-project.org/package=e1071`. R package version 1.6.

M. H. Dominiczak and W. D. Fraser. Blood and plasma proteins. In *Medical Biochemistry and Disease*, pages 31–39. Elsevier Inc., fourth edition, 2014.

R. R. Drake, E. E. Schwegler, G. Malik, J. Diaz, T. Block, A. Mehta, and O. J. Semmes. Lectin capture strategies combined with mass spectrometry for the discovery of serum glycoprotein biomarkers. *Molecular & Cellular Proteomics*, 5 (10):1957–1967, 2006.

P. Du, W. A. Kibbe, and S. M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22:2059–2065, 2006.

W. Dubitzky, M. Granzow, and D. P. Berrar. *Fundamentals of Data Mining in Genomics and Proteomics.* Springer Science and Business Media, 2007.

B. Efron. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394):461–470, 1986.

B. Efron and R. Tibshirani. Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, 1997.

E. A. L. Enninga, W. K. Nevala, D. J. Creedon, S. N. Markovic, and S. G. Holtan. Fetal sex-based differences in maternal hormones, angiogenic factors, and immune mediators during pregnancy and the postpartum period. *American Journal of Reproductive Immunology*, 73(3):251–262, 2015.

W. W. J. Ewens and G. R. Grant. *Statistical methods in bioinformatics: an introduction*. Springer, 2001.

R.-E. Fan, P.-H. Chen, and C.-J. Lin. Working set selection using second order information for training support vector machines. *The Journal of Machine Learning Research*, 6:1889–1918, 2005.

H. Fang and L. Han. Principal component analysis on non-Gaussian dependent data. In *The 30th International Conference on Machine Learning*, volume 28, pages 240–248, Atlanta, USA, June 2013. JMLR Workshop and Conference Proceedings.

T. Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8): 861–874, 2006.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

G. Fleury, A. O. Hero, S. Yoshida, T. Carter, C. Barlow, and A. Swaroop. Pareto analysis for gene filtering in microarray experiments. In *European Signal Processing Conference (EUSIPCO)*, 2002.

J. Fox. *An R and S-Plus Companion to Applied Regression*. Sage Publications, 2002.

E. T. Fung and C. Enderwick. Proteinchip clinical proteomics: computational challenges and solutions. *Biotechniques*, 32(Suppl 1):34–41, 2002.

J. A. Gagnon-Bartsch and T. P. Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.

C. L. Gatlin, K. Y. White, M. B. Tracy, C. E. Wilkins, O. J. Semmes, J. O. Nyalwidhe, R. R. Drake, and D. I. Malyarenko. Enhancement in MALDI-TOF MS analysis of the low molecular weight human serum proteome. *Journal of Mass Spectrometry*, 46(1):85–89, 2011.

A. Gelman. Analysis of varianceâĂŤwhy it is more important than ever. *The Annals of Statistics*, 33(1):1–53, 2005.

T. Gemoll, U. J. Roblick, G. Auer, H. Jornvall, and J. K. Habermann. SELDI-TOF serum proteomics and colorectal cancer: A current overview. *Archives of Physiology and Biochemistry*, 116(4-5):188–196, 2010.

S. Gharbi, P. Gaffney, A. Yang, M. J. Zvelebil, R. Cramer, M. D. Waterfield, and J. F. Timms. Evaluation of two-dimensional differential gel electrophoresis for proteomic expression analysis of a model breast cancer cell system. *Molecular & Cellular Proteomics*, 1(2):91–98, 2002.

S. Gibb and K. Strimmer. MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics*, 28(17):2270–2271, 2012.

J. Gil and M. Werman. Computing 2-d min, median, and max filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:504–507, 1993.

G. L. Glish and R. W. Vachet. The basics of mass spectrometry in the twentyfirst century. *Nature Reviews Drug Discovery*, 2(2):140–150, 2003.

G. H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420, 1970.

L. Gong, W. Constantine, and Y. A. Chen. *msProcess: Protein Mass Spectra Processing*, 2012. URL `http://CRAN.R-project.org/package=msProcess`. R package version 1.0.7.

T. J. Griffin, C. M. Lock, X.-J. Li, A. Patel, I. Chervetsova, H. Lee, M. E. Wright, J. A. Ranish, S. S. Chen, and R. Aebersold. Abundance ratio-dependent proteomic analysis by mass spectrometry. *Analytical chemistry*, 75(4):867–874, 2003.

W. E. Grizzle, A. Bao-Ling, W. L. Bigbee, T. P. Conrads, C. Carroll, F. Ziding, E. Izbicka, M. Jendoubi, D. Johnsey, J. Kagan, R. J. Leach, D. B. McCarthy, O. J. Semmes, S. Srivastava, S. Srivastava, I. M. Thompson, M. D. Thornquist, M. Verma, Z. Zhen, and Z. Zhiqiang. Serum protein expression profiling for cancer detection: Validation of a SELDI-based approach for prostate cancer. *Disease Markers*, 19(4/5):185–195, 2003.

J. Guo. Simultaneous variable selection and class fusion for high-dimensional linear discriminant analysis. *Biostatistics*, 11(4):599–608, 2010.

J. Guo, E. Levina, G. Michailidis, and J. Zhu. Pairwise variable selection for high-dimensional model-based clustering. *Biometrics*, 66(3):793–804, 2010.

Y. Guo, T. Hastie, and R. Tibshirani. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100, 2007.

Y. Guo, T. Hastie, and R. Tibshirani. *rda: Shrunken Centroids Regularized Discriminant Analysis*, 2012. URL `http://CRAN.R-project.org/package=rda`. R package version 1.0.2-2.

D. Hanahan and R. A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, 2011.

L. H. Hartwell, L. Hood, M. L. Goldberg, A. E. Reynolds, L. M. Silver, and R. C. Veres. *Genetics: From Genes to Genomes*. McGraw-Hill, third edition, 2008.

T. Hastie and R. Tibshirani. Classification by pairwise coupling. *The Annals of Statistics*, 26(2):451–471, 1998.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, New York, 2001.

T. Hastie, R. Tibshirani, B. Narasimhan, and G. Chu. *impute: Imputation for microarray data*, 2014. R package version 1.40.0.

D. F. Heitjan and S. Basu. Distinguishing 'missing at random' and 'missing completely at random'. *The American Statistician*, 50(3):207–213, 1996.

C. Heneghan, J. Flynn, M. O'Keefe, and M. Cahill. Characterization of changes in blood vessel width and tortuosity in retinopathy of prematurity using image analysis. *Medical Image Analysis*, 6(4):407–429, 2002.

A. O. Hero and G. Fleury. Pareto-optimal methods for gene ranking. *Journal of VLSI signal processing systems for signal, image and video technology*, 38(3): 259–275, 2004.

G. L. Hortin. The MALDI-TOF mass spectrometric view of the plasma proteome and peptidome. *Clinical Chemistry*, 52(7):1223–1237, 2006.

R. L. Hotz. Here's an omical tale: Scientists discover spreading suffix. *The Wall Street Journal*, August 2012.

L. L. House, M. A. Clyde, and R. L. Wolpert. Nonparametric models for peak identification and quantification in mass spectroscopy, with application to MALDI-TOF. Technical report, Citeseer, 2006.

L. L. House, M. A. Clyde, and R. L. Wolpert. Bayesian nonparametric models for peak identification in MALDI-TOF mass spectroscopy. *The Annals of Applied Statistics*, 5(2B):1488–1511, 2011.

F. Husson and J. Josse. *missMDA: Handling missing values with/in multivariate data analysis (principal component methods)*, 2012. URL `http://CRAN.R-project.org/package=missMDA`. R package version 1.5.

F. Husson, J. Josse, S. Le, and J. Mazet. *FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R*, 2012. URL `http://CRAN.R-project.org/package=FactoMineR`. R package version 1.19.

T. W. Hutchens and T.-T. Yip. New desorption strategies for the mass spectrometric analysis of macromolecules. *Rapid Communications in Mass Spectrometry*, 7(7): 576–580, 1993.

IARC. Australia (2012): estimated cancer mortality, all ages. Technical report, World Health Organisation, International Agency for Research on Cancer, IARC: France, 2015a. URL http://globocan.iarc.fr/old/summary_table_pop-html.asp.

IARC. GLOBOCAN 2012: estimated cancer incidence, mortality and prevalence worldwide in 2012. Technical report, World Health Organisation, International Agency for Research on Cancer, IARC: France, 2015b. URL http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx.

A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research*, 11:1957–2000, 2010.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945, 2004.

M. M. Issa, S. Bux, T. Chun, J. A. Petros, A. J. Labadia, K. Anastasia, L. E. Miller, and F. F. Marshall. A randomized prospective trial of intrarectal lidocaine for pain control during transrectal prostate biopsy: The emory university experience. *The Journal of Urology*, 164(2):397–399, 2000.

H. J. Issaq, T. D. Veenstra, T. P. Conrads, , and D. Felschow. The SELDI-TOF MS approach to proteomics: Protein profiling and biomarker identification. *Biochemical and Biophysical Research Communications*, 292(3):587–592, 2002.

L. Jacob, J. Gagnon-Bartsch, and T. P. Speed. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *ArXiv e-prints*, November 2012. URL http://arxiv.org/abs/1211.4259. Technical report, arXiv, 2012.

B. J. Jenkins, D. Grail1, T. Nheu, M. Najdovska1, B. Wang, P. Waring, M. Inglese, R. M. McLoughlin, S. A. Jones, N. Topley, H. Baumann, L. M. Judd, A. S. Giraud, A. Boussioutas, H.-J. Zhu, and M. Ernst. Hyperactivation of Stat3 in gp130 mutant mice promotes gastric hyperproliferation and desensitizes TGF-signaling. *Nature Medicine*, 11(0):845–852, 2005.

O. N. Jensen. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Current opinion in chemical biology*, 8(1):33–41, 2004.

P. Johansson and M. Ringner. Classification of genomic and proteomic data using support vector machines. In W. Dubitzky, M. Granzow, and D. Berrar, editors, *Fundamentals of Data Mining in Genomics and Proteomics*, pages 187–202. Springer, 2004.

I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2005.

L. M. Judd, M. Ulaganathan, M. Howlett, and A. S. Giraud. Cytokine signalling by gp130 regulates gastric mucosal healing after ulceration and, indirectly, antral tumour progression. *The Journal of Pathology*, 217(4):552–562, 2009.

A. Karatzoglou, D. Meyer, and K. Hornik. Support vector machines in R. *Journal of Statistical Software*, 15(9), 2005.

I. D. Karbassi, J. O. Nyalwidhe, C. E. Wilkins, L. H. Cazares, R. S. Lance, O. J. Semmes, and R. R. Drake. Proteomic expression profiling and identification of serum proteins using immobilized trypsin beads with MALDI-TOF/TOF. *Journal of Proteome Research*, 8(9):4182–4192, 2009.

Y. V. Karpievitch, T. Taverner, J. N. Adkins, S. J. Callister, G. A. Anderson, R. D. Smith, and A. R. Dabney. Normalization of peak intensities in bottom-up ms-based proteomics using singular value decomposition. *Bioinformatics*, 25(19): 2573–2580, 2009.

Y. V. Karpievitch, A. R. Dabney, and R. D. Smith. Normalization and missing value imputation for label-free LC-MS analysis. *BMC bioinformatics*, 13(Suppl 16):S5, 2012.

S. A Kazmi, S. Ghosh, D.-G. Shin, D. W. Hill, and D. F. Grant. Alignment of high resolution mass spectra: development of a heuristic approach for metabolomics. *Metabolomics*, 2(2):75–83, 2006.

S. Kim, I. Koo, A. Fang, and X. Zhang. Smith-Waterman peak alignment for comprehensive two-dimensional gas chromatography-mass spectrometry. *BMC bioinformatics*, 12(1):235, 2011.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14:2, pages 1137–1145, 1995.

J. J. Kurinczuk, D. E. Parsons, V. Dawes, and P. R. Burton. The relationship between asthma and smoking during pregnancy. *Women & Health*, 29(3):31–47, 1999.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

J. Lederberg and A. T. McCray. ‘Ome Sweet ‘Omics– A genealogical treasury of words. *The Scientist*, 15(7):8, April 2001.

M.-L. T. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences*, 97(18):9834–9839, 2000.

J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):1724–1735, 2007.

M. Li, S. Chen, J. Zhang, H. Chen, and Y. Shyr. Wave-spec: a preprocessing package for mass spectrometry data. *Bioinformatics*, 27(5):739–740, 2011.

X. Li. *PROcess: Ciphergen SELDI-TOF Processing*, 2005. URL `http://bioconductor.org/packages/release/bioc/html/PROcess.html`. R package version 1.42.0.

K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

A. Liaw and M. Wiener. Classification and regression by randomForest. *R News*, 2 (3):18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

G. Litwack. Blood and lymphatic system. In *Human Biochemistry and Disease*, pages 853–898. Elsevier Inc., 2008.

G. Malik, M. D. Ward, S. K. Gupta, M. W. Trosset, W. E. Grizzle, B.-L. Adam, J. I. Diaz, and O. J. Semmes. Serum levels of an isoform of apolipoprotein A-II as a potential marker for prostate cancer. *Clinical Cancer Research*, 11(3):1073–1085, 2005.

D. Mantini, F. Petrucci, D. Pieragostino, P. D. Boccio, P. Sacchetta, G. Candiano, G. M. Ghiggeri, A. Lugaresi, G. Federici, C. D. Ilio, and A. Urbani. A computational platform for MALDI-TOF mass spectrometry data: Application to serum and plasma samples. *Journal of Proteomics*, 73(3):562–570, 2010.

C.-D. Mayer and C. A. Glasbey. Statistical methods in microarray gene expression data analysis. In D. Husmeier, R. Dybowski, and S. Roberts, editors, *Probabilistic Modeling in Bioinformatics and Medical Informatics*, Advanced Information and Knowledge Processing, pages 211–238. Springer London, 2005.

R. A. McLean, W. L. Sanders, and W. W. Stroup. A unified approach to mixed linear models. *The American Statistician*, 45(1):54–64, 1991.

D. McLerran, W. E. Grizzle, Z. Feng, W. L. Bigbee, L. L. Banez, L. H. Cazares, D. W. Chan, J. Diaz, E. Izbicka, J. Kagan, D. E. Malehorn, G. Malik, D. Oelschlager, A. Partin, T. Randolph, N. Rosenzweig, S. Srivastava, S. Srivastava, I. M. Thompson, M. Thornquist, D. Troyer, Y. Yasui, Z. Zhang, L. Zhu, and O. J. Semmes. Analytical validation of serum proteomic profiling for diagnosis of prostate cancer: Sources of sample bias. *Clinical Chemistry*, 54(1):44–52, 2008a.

D. McLerran, W. E. Grizzle, Z. Feng, I. M. Thompson, W. L. Bigbee, L. H. Cazares, D. W. Chan, J. Dahlgren, J. Diaz, J. Kagan, D. W. Lin, G. Malik, D. Oelschlager, A. Partin, T. W. Randolph, L. Sokoll, S. Srivastava, S. Srivastava, M. Thornquist, D. Troyer, G. L. Wright, Z. Zhang, L. Zhu, and O. J. Semmes. SELDI-TOF MS whole serum proteomic profiling with IMAC surface does not reliably detect prostate cancer. *Clinical Chemistry*, 54(1):53–60, 2008b.

S. Meding, U. Nitsche, B. Balluff, M. Elsner, S. Rauser, C. Schöne, M. Nipp, M. Maak, M. Feith, M. P. Ebert, H. Friess, R. Langer, H. Höfler, H. Zitzelsberger, R. Rosenberg, and A. Walch. Tumor classification of six common cancer types based on proteomic profiling by MALDI imaging. *Journal of Proteome Research*, 11(3):1996–2003, 2012.

M. Merchant and S. R. Weinberger. Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry. *Electrophoresis*, 21(6): 1164–1177, 2000.

W. Meuleman, J. Engwegen, M.-C. Gast, J. Beijnen, M. Reinders, and L. Wessels. Comparison of normalisation methods for surface-enhanced laser desorption and ionisation (SELDI) time-of-flight (TOF) mass spectrometry data. *BMC Bioinformatics*, 9(1):88, 2008.

B. L. Mitchell, Y. Yasui, J. W. Lampe, P. R. Gafken, and P. D. Lampe. Evaluation of matrix-assisted laser desorption/ionization-time of flight mass spectrometry proteomic profiling: identification of alpha 2-HS glycoprotein B-chain as a biomarker of diet. *Proteomics*, 5(8):2238–2246, 2005.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. The MIT Press, 2012.

J. S. Morris, K. R. Coombes, J. Koomen, K. A. Baggerly, and R. Kobayashi. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21(9):1764–1775, 2005.

V. A. Moyer. Screening for prostate cancer: U.S. preventive services task force recommendation statement. *Annals of Internal Medicine*, 157(2):120–134, 2012.

V. E. Murphy, T. Zakar, R. Smith, W. B. Giles, P. G. Gibson, and V. L. Clifton. Reduced 11β-hydroxysteroid dehydrogenase type 2 activity is associated with decreased birth weight centile in pregnancies complicated by asthma. *Journal of Clinical Endocrinology & Metabolism*, 87(4):1660–1668, 2002.

V. E. Murphy, R. F. Johnson, Y.-C. Wang, K. Akinsanya, P. G. Gibson, R. Smith, and V. L. Clifton. The effect of maternal asthma on placental and cord blood protein profiles. *Journal of the Society for Gynecologic Investigation*, 12(5):349–355, 2005.

V. E. Murphy, R. F. Johnson, Y.-C. Wang, K. Akinsanya, P. G. Gibson, R. Smith, and V. L. Clifton. Proteomic study of plasma proteins in pregnant women with asthma. *Respirology*, 11:41–48, 2006.

S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A*, 135(3):370–384, 1972.

W. M. Old, K. Meyer-Arendt, L. Aveline-Wolf, K. G. Pierce, A. Mendoza, J. R. Sevinsky, K. A. Resing, and N. G. Ahn. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Molecular & Cellular Proteomics*, 4(10):1487–1502, 2005.

G. S. Omenn, D. J. States, M. Adamski, T. W. Blackwell, R. Menon, H. Hermjakob, R. Apweiler, B. B. Haab, R. J. Simpson, J. S. Eddes, E. A. Kapp, R. L. Moritz, D. W. Chan, A. J. Rai, A. Admon, R. Aebersold, J. Eng, W. S. Hancock, S. A. Hefta, H. Meyer, Y.-K. Paik, J.-S. Yoo, P. Ping, J. Pounds, J. Adkins, X. Qian, R. Wang, V. Wasinger, C. Y. Wu, X. Zhao, R. Zeng, A. Archakov, A. Tsugita, I. Beer, A. Pandey, M. Pisano, P. Andrews, H. Tammen, D. W. Speicher, and S. M. Hanash. Overview of the HUPO plasma proteome project: Results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics*, 5(13):3226–3245, 2005.

S. J. Orfanidis. *Introduction to Signal Processing.* Prentice Hall Signal Processing Series. Prentice Hall, 1996.

E. Orvisky, S. K. Drake, B. M. Martin, M. Abdel-Hamid, H. W. Ressom, R. S. Varghese, Y. An, D. Saha, G. L. Hortin, and C. A. Loffredo. Enrichment of low molecular weight fraction of serum for ms analysis of peptides associated with hepatocellular carcinoma. *Proteomics*, 6(9):2895–2902, 2006.

Q. Pan, L. W. Bao, C. G. Kleer, M. S. Sabel, K. A. Griffith, T. N. Teknos, and S. D. Merajver. Protein kinase Cε is a predictive biomarker of aggressive breast cancer and a validated target for RNA interference anticancer therapy. *Cancer Research*, 65(18):8366–8371, 2005.

S. Parpart, A. Rudis, A. Schreck, N. Dewan, and P. Warren. Sensitivity and specificity in prostate cancer screening methods and strategies. *The Journal of Young Investigators*, 16(4), 2007.

K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

E. Pennisi. ENCODE project writes eulogy for junk DNA. *Science*, 337(6099): 1159–1161, 2012.

M. A. S. Penno, M. Klingler-Hoffmann, J. A. Brazzatti, A. Boussioutas, T. Putoczki, M. Ernst, and P. Hoffmann. 2D-DIGE analysis of sera from transgenic mouse models reveals novel candidate protein biomarkers for human gastric cancer. *Journal of Proteomics*, 77(0):40–58, 2012.

E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359(9306): 572–577, 2002a.

E. F. Petricoin, D. K. Ornstein, C. P. Paweletz, A. Ardekani, P. S. Hackett, B. A. Hitt, A. Velassco, C. Trucco, L. Wiegand, K. Wood, C. B. Simone, P.J. Levine, W. M. Linehan, M. R. Emmert-Buck, S. M. Steinberg, E. C. Kohn, and L. A. Liotta. Serum proteomic patterns for detection of prostate cancer. *Journal of the National Cancer Institute*, 94(20):1576–1578, 2002b.

J.C. Pinheiro and D.M. Bates. *Mixed-Effects Models in S and S-Plus*. Statistics and Computing. Springer, 2009.

M. J. D. Powell. The BOBYQA algorithm for bound constrained optimization without derivatives. Technical report, Centre for Mathematical Sciences, University of Cambridge, August 2009. DAMTP2009/NA06.

J. R. Prensner, M. A. Rubin, J. T. Wei, and A. M. Chinnaiyan. Beyond PSA: The next generation of prostate cancer biomarkers. *Science Translational Medicine*, 4 (127), 2012.

W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipes in FORTRAN 77 Vol. 1: The Art of Scientific Computing*. Fortran Numerical Recipes. Cambridge University Press, 1992.

W. H. Pun. Multilevel and Bayesian models for the analysis of mass spectrometry data. Honours thesis, The University of Adelaide, 2014.

Y. Qu, B.-L. Adam, Y. Yasui, M. D. Ward, L. H. Cazares, P. F. Schellhammer, Z. Feng, O. J. Semmes, and G. L. Wright. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clinical Chemistry*, 48(10):1835–1843, 2002.

J. Quackenbush. Microarray data normalization and transformation. *Nature genetics*, 32:496–501, 2002.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

T. W. Randolph. Scale-based normalization of spectral data. *Cancer Biomarkers*, 2(3):135–144, 2006.

H. W. Ressom, R. S. Varghese, M. Abdel-Hamid, S. A.-L. Eissa, D. Saha, L. Goldman, E. F. Petricoin, T. P. Conrads, T. D. Veenstra, C. A. Loffredo, and R. Goldman. Analysis of mass spectral serum profiles for biomarker selection. *Bioinformatics*, 21(21):4039–4045, 2005.

D. P. Rice, T. A. Hodgson, and A. N. Kopstein. The economic costs of illness: a replication and update. *Health care financing review*, 7(1):61, 1985.

G. K. Robinson. That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6(1):15–32, 1991.

M. D. Robinson. *Methods for the analysis of gas chromatography - mass spectrometry data*, 2008. PhD dissertation, University of Melbourne.

M. D. Robinson. *flagme: Analysis of Metabolomics GC/MS Data*, 2010. R package version 1.12.0.

M. D. Robinson, D. P. De Souza, W. W. Keen, E. C. Saunders, M. J. McConville, T. P. Speed, and V. A. Likic. A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments. *BMC Bioinformatics*, 8(1):419, 2007.

G. J. M. Rosa, J. P. Steibel, and R. J. Tempelman. Reassessing design and analysis of two-colour microarray experiments using mixed effects models. *Comparative and Functional Genomics*, 6(3):123–131, 2005.

P. Roy, C. Truntzer, D. Maucort-Boulch, T. Jouve, and N. Molinari. Protein mass spectra data analysis for clinical biomarker discovery: a global review. *Briefings in bioinformatics*, 12(2):176–186, 2011.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

R. W. Ruddon. *Cancer Biology*. Oxford University Press, USA, 2007.

P. A. Rudnick, K. R. Clauser, L. E. Kilpatrick, D. V. Tchekhovskoi, P. Neta, N. Blonder, D. D. Billheimer, and R. K. Blackman. Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Molecular & Cellular Proteomics*, 9(2):225–241, 2010.

C. Saha and M. P. Jones. Asymptotic bias in the linear mixed effects model under non-ignorable missing data mechanisms. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(1):167–182, 2005.

A. C. Sauve and T. P. Speed. Normalization, baseline correction and alignment of high-throughput mass spectrometry data. *Proceedings of the Genomic Signal Processing and Statistics Workshop*, 2004.

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

F. H. Schroder, J. Hugosson, M. J. Roobol, T. L. J. Tammela, S. Ciatto, V. Nelen, M. Kwiatkowski, M. Lujan, H. Lilja, M. Zappa, L. J. Denis, F. Recker, A. Berenguer, L. Maattanen, C. H. Bangma, G. Aus, A. Villers, X. Rebillard, T. van der Kwast, B. G. Blijenberg, S. M. Moss, H. J. de Koning, and A. Auvinen. Screening and prostate-cancer mortality in a randomized european study. *The New England Journal of Medicine*, 360(13):1320–1328, 2009.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.

O. J. Semmes, Z. Feng, B.-L. Adam, L. L. Banez, W. L. Bigbee, D. Campos, L. H. Cazares, D. W. Chan, W. E. Grizzle, E. Izbicka, J. Kagan, G. Malik, D. McLerran, J. W. Moul, A. Partin, P. Prasanna, J. Rosenzweig, L. J. Sokoll, S. Srivastava, S. Srivastava, I. Thompson, M. J. Welsh, N. White, M. Winget, Y. Yasui, Z. Zhang, and L. Zhu. Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. assessment of platform reproducibility. *Clinical Chemistry*, 51 (1):102–112, 2005.

Y. Shi, R. Xiang, C. Hovárth, and J. A. Wilkins. The role of liquid chromatography in proteomics. *Journal of Chromatography A*, 1053(1-2):27–36, 2004.

J. Shlens. A tutorial on principal component analysis. Technical report, Center for Neural Science, New York University, April 2009. Version 3.01.

G. Siuzdak. *The Expanding Role of Mass Spectrometry in Biotechnology*. McC Press, 2006.

C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification. *Analytical Chemistry*, 78:779–787, 2006.

T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

G. K. Smyth. limma: Linear models for microarray data. In R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Statistics for Biology and Health, pages 397–420. Springer New York, 2005.

G. K. Smyth and T. Speed. Normalization of cDNA microarray data. *Methods*, 31 (4):265–273, 2003.

T. A. B. Snijders and R. J. Bosker. *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Sage Publishers, London, second edition, 2012.

I. Soerjomataram, J. Lortet-Tieulent, D. M. Parkin, J. Ferlay, C. Mathers, D. Forman, and F. Bray. Global burden of cancer in 2008: a systematic analysis of disability-adjusted life-years in 12 world regions. *The Lancet*, 380(9856):1840–1850, 2012.

P. Soille. *Morphological Image Analysis: Principles and Applications.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1999.

P. J. Solomon. Some statistics in bioinformatics: The fifth Armitage Lecture. *Statistics in Medicine*, 28(23):2833–2856, 2009.

J. M. Sorace and M. Zhan. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics*, 4(24), 2003.

V. Swarup and M. R. Rajeswari. Circulating (cell-free) nucleic acids–a promising, non-invasive tool for early detection of several human diseases. *FEBS letters*, 581 (5):795–799, 2007.

T. Taverner. *DanteR: Proteomics data analysis tool*, 2012. R package version 0.2.

N. C. Tebbutt, A. S. Giraud, M. Inglese, B. Jenkins, P. Waring, F. J. Clay, S. Malki, B. M. Alderman, D. Graill1, F. Hollande, J. K. Heath, and M. Ernst. Reciprocal regulation of gastrointestinal homeostasis by SHP2 and STAT-mediated trefoil gene activation in gp130 mutant mice. *Nature Medicine*, 8(0):1089–1097, 2002.

R. Terracciano, L. Pasqua, F. Casadonte, S. Frasca, M. Preiano, D. Falcone, and R. Savino. Derivatized mesoporous silica beads for MALDI-TOF MS profiling of human plasma and urine. *Bioconjugate Chemistry*, 20(5):913–923, 2009.

The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, 2015.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B Methodological*, 58(1):267–288, 1996.

R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, pages 104–117, 2003.

R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, and Q.-T. Le. Sample classification from protein mass spectrometry, by ‘peak probability contrasts’. *Bioinformatics*, 20(17):3034–3044, 2004.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

UniProt, Accessed February 2013. `http://www.uniprot.org/`.

M. van Herk. A fast algorithm for local minimum and maximum filters on rectangular and octagonal kernels. *Pattern Recognition Letters*, 13(7):517–521, 1992.

M. van Herk, J. C. de Munck, J. V. Lebesque, S. Muller, C. Rasch, and A. Touw. Automatic registration of pelvic computed tomography data and magnetic resonance scans including a full circle method for quantitative accuracy evaluation. *Medical physics*, 25:2054, 1998.

V. N Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics With S*. Statistics and Computing. Springer, 2002.

J. Villanueva, D. R. Shaffer, J. Philip, C. A. Chaparro, H. Erdjument-Bromage, A. B. Olshen, M. Fleisher, H. Lilja, E. Brogi, J. Boyd, M. Sanchez-Carbayo, E. C. Holland, C. Cordon-Cardo, H. I. Scher, and P. Tempst. Differential exoprotease activities confer tumor-specific serum peptidome patterns. *The Journal of Clinical Investigation*, 116(1):271–284, 2006.

J. A. Vizcaino, R. G. Côté, A. Csordas, J. A. Dianes, A. Fabregat, J. M. Foster, J. Griss, E. Alpi, M. Birim, J. Contell, G. O’Kelly, A. Schoenegger, D. Ovelleiro, Y. Pérez-Riverol, F. Reisinger, D. Ríos, R. Wang, and H. Hermjakob. The proteomics identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Research*, 41:1063–1069, 2013.

C. Walsh. *Posttranslational Modification Of Proteins: Expanding Nature’s Inventory*. Roberts, 2006.

Y.-G. Wang and X. Lin. Effects of variance-function misspecification in analysis of longitudinal data. *Biometrics*, 61(2):413–421, 2005.

B. T. West, K. B. Welch, and A. T. Galecki. *Linear mixed models: a practical guide using statistical software.* CRC Press, first edition, 2007.

W. C. Wiley and L. H. McLaren. Time-of-flight mass spectrometer with improved resolution. *The Review of Scientific Instruments*, 26(12):1150–1157, 1955.

R. D. Wolfinger, G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. S. Paules. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, 8(6):625–637, 2001.

W. Wolski, M. Lalowski, P. Martus, R. Herwig, P. Giavalisco, J. Gobom, A. Sickmann, H. Lehrach, and K. Reinert. Transformation and other factors of the peptide mass spectrometry pairwise peak-list comparison process. *BMC Bioinformatics*, 6(1):285, 2005.

J. W. H. Wong, G. Cagney, and H. M. Cartwright. SpecAlign: processing and alignment of mass spectra datasets. *Bioinformatics*, 21(9):2088–2090, 2005.

B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19(13):1636–1643, 2003.

C. Yang, Z. He, and W. Yu. Comparison of public peak detection algorithms for maldi mass spectrometry data analysis. *BMC Bioinformatics*, 10(1), 2009.

Y. H. Yang and T. Speed. Design issues for cDNA microarray experiments. *Nature Reviews Genetics*, 3(1):579–588, 2002.

Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*, 11:108–136, 2002.

Y. Yasui, D. McLerran, B-L. Adam, M. Winget, M. Thornquist, and Ziding Feng. An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers. *Journal of Biomedicine and Biotechnology*, 2003(4):242–248, 2003.

J. R. Yates III. A century of mass spectrometry: from atoms to proteomes. *Nature Methods*, 8:633–637, 2011.

S. L. Zeger and K.-Y. Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, pages 121–130, 1986.

G. Zhang, B. M. Ueberheide, S. Waldemarson, S. Myung, K. Molloy, J. Eriksson, B. T. Chait, T. A. Neubert, and David Fenyö. Protein quantitation using mass spectrometry. In *Computational Biology*, pages 211–222. Springer, 2010.