Statistical Analysis of Proteomic Mass Spectrometry Data

by Kelly Handley, BSc(Hons)
Division of Statistics

School of Mathematical Sciences
University of Nottingham

Thesis submitted to the University of Nottingham
for the Degree of Doctor of Philosophy,

March 2007

Abstract

This thesis considers the statistical modelling and analysis of proteomic mass spectrometry data. Proteomics is a relatively new field of study and tried and tested methods of analysis do not yet exist. Mass spectrometry output is high-dimensional and so we firstly develop an algorithm to identify peaks in the spectra in order to reduce the dimensionality of the datasets. We use the results along with a variety of classification methods to examine the classification of new spectra based on a training set. Another method to reduce the complexity of the problem is to fit a parametric model to the data. We model the data as a mixture of Gaussian peaks with parameters representing the peak locations, heights and variances, and apply a Bayesian Markov chain Monte Carlo (MCMC) algorithm to obtain their estimates. These resulting estimates are used to identify m/z values where differences are apparent between groups, where the m/z value of an ion is its mass divided by its charge. A multilevel modelling framework is also considered to incorporate the structure in the data and locations exhibiting differences are again obtained.

We consider two mass spectrometry datasets in detail. The first consists of mass spectra from breast cancer cells which either have or have not been treated with the chemotherapeutic agent Taxol. The second consists of mass spectra from melanoma cells classified as stage I or stage IV using the TNM system. Using the MCMC and multilevel techniques described above we show that, in both datasets, small subsets of the available m/z values can be identified which exhibit significant differences in protein expression between groups. Also we see that good classification of new data can also be achieved using a small number of m/z values and that the classification rate does not fall greatly when compared with results from the complete spectra. For both datasets we compare our results with those in the literature which use other techniques on the same data. We conclude by discussing potential areas for further research.

Acknowledgments

Firstly I am eternally grateful to my two supervisors Ian Dryden and William Browne. I would like to thank them for their advice, for proof reading my work and telling me when I had written utter rubbish, and for making me feel left out when other students complained about their supervisors. Also thanks to my examiners, Malcolm Farrow and Andrew Wood, and all the other staff who have taught me over the past six years or so - your knowledge and advice has been invaluable.

I would like to thank all of my maths friends who have kept me just about sane during this PhD. They should be commended for ensuring that no tea break has passed without laughter, excessive geekiness, and of course the perennial question of who would win in a fight between a polar bear and a hippo ...or a badger and a swan ...or Mr Bump and Mr Strong .... Special mentions go to Laura Hobley and also to my fabulous office mates Chris Brignell, Simon Spencer and Kim Evans. They were always there to cheer me up when things didn't work and have put up with many a moan about C++ segmentation faults.

Carolyn, Fiona and my resident tutor colleagues from Newark Hall should also be mentioned. They have been a great group of people to work with for four years and are even nice at 3am when students set the fire alarm off. They also helped me survive the PhD experience by talking to me about things other than maths.

My family should be especially thanked for supporting me both during my first degree and this one. To mom and dad, thanks for your love and support and for putting up with me for longer than anybody else. To Kay, thanks for keeping me suitably amused with humorous emails during my studies and all manner of other things.

I would also like to acknowledge Andrew Smith and Anne Jones. They helped me realise that maths was definitely for me and thus are partly to blame for me being a student for so long!

Finally I am grateful to the University of Nottingham for allowing me to study here, to the Engineering and Physical Sciences Research Council for their financial assistance, to Shahid Mian for the datasets and helpful discussions, to Dave Parkin for the many occasions he fixed my computer when I did silly things to it, to the wonderful maths secretaries who gave us free cake and to the nice people who invented LaTeX. Without them this thesis would never have been written.

> *I've put my heart and soul into this work - and lost my mind in the process.*
>
> Vincent Van Gogh

Contents

List of Figures

List of Tables

Chapter 1

Thesis Outline and Background Information

1.1 Thesis Outline

The research presented in this thesis is concerned with the analysis of proteomic mass spectrometry data. Proteomics, the study of proteins and their functions, is a relatively new field of research and thus there exist no tried and tested methods of analysis.

In Chapter 2 we consider the problem of how to classify new spectra. Proteomic datasets are generally high-dimensional and contain a relatively small number of spectra. Having more observations per spectra than actual spectra could result in perfect classifiers being obtained for the current data which do not generalise to larger datasets. To reduce this problem, a deterministic peak finding algorithm is developed in order to obtain a computationally cheap approximation of the data. The results of the algorithm allow some test spectra to be classified based on a set of training spectra. Small groups of m/z values are identified which provide high rates of correct classification.

In Chapters 3 and 4 we consider the problem of creating a parametric model for the mass spectra. Parametric modelling is another method by which we can reduce the dimension of the problem. In chapter 3 Markov Chain Monte Carlo (MCMC) methods are employed to model the datasets. A number of related models, increasing in complexity, are considered and analysed. The peak finding algorithm developed in chapter 2 is incorporated to provide a suitable starting point for the MCMC algorithm. The resulting parameter values are used to identify differences between different mass spectra which are then compared with results obtained by other researchers. Some of this research is published in Handley et al., (2005)

In chapter 4 we consider modelling the data using multilevel models. The data have a natural hierarchical structure (m/z values within spectra) and research in previous chapters has ignored this information. The peak finding procedure from chapter 2 is again implemented to assist in the construction of the fixed and random effects. For each group we fit a series of fixed effects which model the average spectra for that group. Random effects are also incorporated in order that each spectrum can be modelled instead of simply the average for each group. Differences between the groups will be identified by considering the differences between fixed effects. A selection of m/z values are obtained for each of the datasets which indicate where groups differ.

Lastly, in Chapter 5 we conclude by discussing the main results drawn from this research. Also possible areas for further work are presented.

The results and graphics in this thesis have been obtained using the C++ programming language (Stroustrup, 2003) and the software packages R (R Development Core Team, 2005) and MLwiN (Rasbash et al., 2000). Time comparisons were calculated by using the high performance GRID computer.

1.2 Background Information

Due to the high mortality rate of patients with advanced forms of the disease, the detection of early stage cancer is of great importance. The identification of indicators (biomarkers) which enable the progress of a disease to be measured is the subject of much research as there is the potential to diagnose patients before they present with symptoms. Biomarkers assist in identifying disease progression by providing quantitative information about molecules at any given point in time. If the relative concentration of a particular molecule is found to be different between healthy and diseased cases there is the potential to use this information to diagnose disease. The identification of biomarkers suitable for the early detection and diagnosis of cancer could greatly improve patients' diagnoses. Early detection of cancer could result in less severe, more treatable diseases and ultimately higher cure rates.

Shown in table 1.1 are some examples of biomarkers already being used for disease diagnosis (Diamandis, 2004). However, none of these biomarkers are suitable for general cancer screening as they are not sufficiently specific and lead to many false-positive results. Some of the procedures that are currently used for general screening are mammography for breast cancer, pap smear for cervical cancer and colonoscopy for colorectal cancer (Tibshirani et al., 2004).

Microarray technology is a popular method of expressing an entire genome on a single chip by analysing mRNA. mRNA is genetic material created in the nucleus of a cell. It is a copy of a small section of DNA (a gene) and contains information which tells the cell how to synthesise proteins. The expression levels of many thousands of genes can be simultaneously studied and differences in relative expressions can be used to infer differing disease states. However, proteins are considered closer to actual biological functions than mRNA and, since mRNA concentration correlates poorly with protein concentration (Yasui et al., 2003), it is more useful to look for protein biomarkers for disease. This is not possible using microarrays and so the need for large scale analy-

biomarker	cancer type
α-Fetoprotein (AFP)	hepatoma; testicular
Carcinoembryonic antigen (CEA)	colon; breast; lung; pancreatic
Prostate specific antigen (PSA)	prostate
CA125	ovarian
CA15.3	breast
CA19.9	gastrointestinal
Immunoglobins	B cell dyscrasias
Chroriogonadotrophin	testicular; trophoblastic tumours
Steroid hormone receptors	breast

Table 1.1: Examples of established cancer biomarkers.

sis of proteins led to the new research area of proteomics which is concerned with characterising all the proteins in a biological sample. Nearly all useful biomarkers identified thus far have been proteins (Robbins et al., 2005).

The traditional method for discovering disease-associated proteins was by using two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), invented by O'Farrell and Klose in 1975 (O'Farrell, 1975 and Klose, 1975). This method separates proteins onto a gel by pH level and size, and gels can then be compared between two or more disease groups. The large number of studies carried out using this method has enabled databases to be created of disease-associated proteins. Some studies that have used this method are, for example, Edwards et al. (1982), Adam et al. (2001) and Srinivas et al. (2001) which between them cover prostate and bladder cancers. Although 2D-PAGE is able to resolve thousands of proteins and detect differences in protein expression, it is labour intensive and requires high abundance of the proteins in question. Also some types of proteins - hydrophobic, strongly acidic or strongly basic - are poorly resolved (Poon et al., 2003). Nearly all common cancer biomarkers have a concentration of 1 ng/mL or less which is below the detection limit for 2D-PAGE and thus no new biomarkers for cancer arose from this method (Robbins et al., 2005).

An advance in the study of proteins was made in 1987 with the development of matrix assisted laser desorption/ionisation time-of-flight (MALDI-TOF) mass spectrometry. Mass spectrometry is a technique used to identify and measure a wide variety of biological and chemical compounds. It works on the principle that different molecules have different masses and thus by separating a substance into its constituents according to their mass we can identify which molecules are present. Mass to charge ratios of up to 50,000 Da can be resolved using this method (Yanagisawa et al., 2003).

Surface enhanced laser desorption/ionisation time-of-flight (SELDI-TOF) mass spectrometry, originally described by Hutchens and Yip (1993), is a novel development in time-of-flight mass spectrometry. The general principle behind SELDI-TOF is that proteins of interest from biological samples bind selectively to a chemical surface, and the impurities are then washed away. This remaining part of the sample is then complexed with an energy absorbing molecule, and analysed by laser desorption/ionisation time-of-flight mass spectrometry to determine the abundance of the different molecules present in the sample. The results are often displayed as a graph showing the relative abundance associated with protein mass/charge (m/z) ratios over a particular Dalton range (see figure 1.1). The main advantage of TOF methods over 2D-PAGE is their ability to detect molecules with m/z ratios smaller than 20 kDa (Qu et al., 2002).

Many different methods have been employed in recent years to analyse mass spectrometry data. Genetic algorithms are used by Petricoin et al. (2002) to study ovarian cancer and artificial neural networks (ANN) by Zhang et al. (1999) and Mian et al. (2003) to study pelvic masses and breast cancer respectively. T-statistics are quite popular and are used, for example, by Chen et al. (2002), Vlahou et al. (2001) and Xiao et al. (2001) to study lung adenocarcinoma, bladder cancer and prostate cancer respectively. Also random forests are used by Izmirlian (2004) and decision trees by Adam et al. (2002)

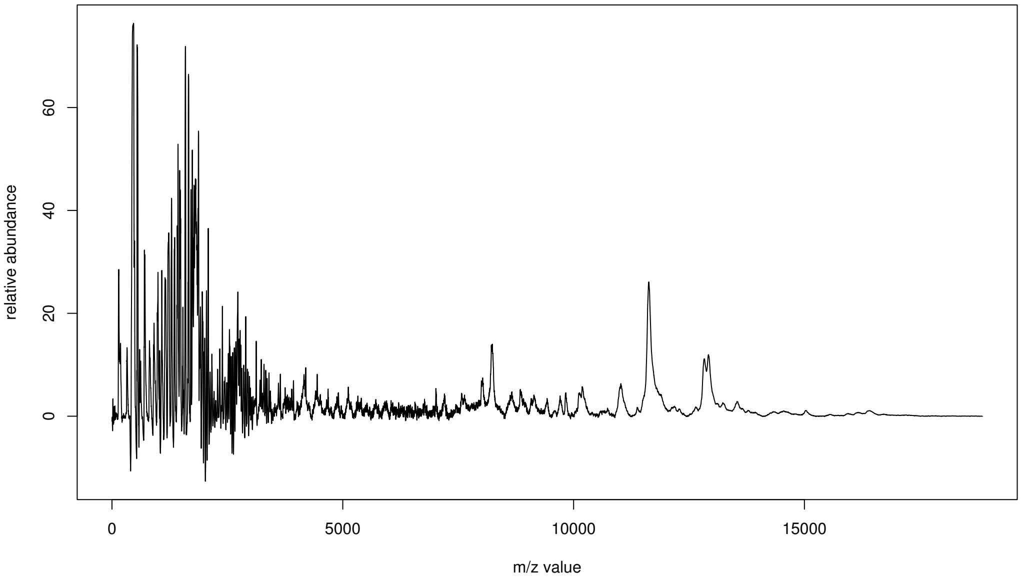

Figure 1.1: An example of a SELDI-TOF mass spectrum.

to study prostate cancer.

The SELDI-TOF method has also been used in some non-cancer settings. Nomura et al. (2004) use the method to search for biomarkers for alcoholism. Uchida et al (2002) study a dataset on rheumatoid arthritis and actually identify one of the biomarkers they find in their analysis as a protein known to be related to the disease. This shows promise for the future of identifying biomarkers by this method.

However, questions have been asked regarding the impressive results quoted in the literature obtained using these proteomic methods. There appear to be four main objections. Firstly, Baggerly et al. (2004) question the reproducibility of the results. They analyse the same data as Petricoin et al. (2002) but are unable to replicate most of their findings as two of the datasets have been *back-*

ground subtracted. This is a process by which electronic noise inherent in the mass spectrometer is removed from the resulting mass spectrum. Background subtraction is an irreversible, non-linear operation and thus the original values cannot be reconstructed. Note that in figure 1.1 there are some negative intensities near the beginning of the spectrum. This is due to the data having been background subtracted. It is also suggested that changes in the laboratory, the mass spectrometer or the ProteinChip array could alter the results obtained.

A recent study by Munro et al. (2006) is the first to have demonstrated the reproducibility of their results. A random forest classifier was used to predict the presence of transitional cell carcinoma (TCC) and initial results indicated 71.7% sensitivity and 62.5% specificity. When an independent validation set was studied 6 months later the respective results were 78.3% and 65.0% which are comparable with tests currently being used to diagnose TCC.

Secondly, both Baggerly et al. (2004) and Sorace and Zhan (2003) question the use of m/z values below 2,000 Daltons. Such m/z values are generally considered to be noise and should be excluded from any analysis. In the paper by Petricoin et al. (2002) classifiers solely in the noise region are obtained that classify the data perfectly. One particular example is found with an m/z value of 2.79 Daltons suggesting an experimental bias not related to disease state.

Thirdly, Banks et al. (2005) conclude that sample handling can have a marked effect on the spectra obtained from time-of-flight mass spectrometry. The time elapsed between obtaining the sample and processing it was a main cause of differences although others included temperature, storage and centrifugation (time and speed).

Lastly, Somorjai et al. (2003) believe that the near-perfect classification rates quoted in the literature are misleading and believe this is due to two reasons - too many m/z values and not enough spectra. With such a large number of m/z values corrections must be made for multiple measurements using, for example, Benjamini and Hochberg's (1995, 2000) false discovery rate approach. Also seemingly robust classifiers can be obtained relatively easily if there are only a few examples per group available. Future experiments should obtain as many examples in each group as is practical to reduce this problem.

1.3 The Datasets

The two datasets analysed in this thesis result from mass spectrometric analysis of breast cancer and melanoma cell lines respectively.

A cell-line consists of cells of a single type taken from an animal or human and grown in the laboratory. These cells can grow and replicate continuously outside the living organism and, with the proper conditions, may be kept alive indefinitely in a Petri dish. All the cells are genetically identical to a single common ancestor cell which makes them valuable for research.

Breast cancer is cancer of breast tissue and is the most common form of cancer in females. In the Western world it will affect approximately one in nine women at some stage of their life (Markham, 2005).

The breast cancer dataset consists of 144 cell-lines divided equally into one of three groups. The first group consists of 48 cell-lines of the type MCF-7/ADR which are chemoresistant and the second and third groups each consist of 48 chemosensitive cell-lines of the types T47D and MCF-7 respectively. Half of the cell-lines in each group have received a 24 hour exposure to the chemotherapeutic agent Taxol.

Taxol (Paclitaxel) is a drug used in the treatment of cancer. It was isolated by Drs. M.E. Wall and M.C. Wani in 1967 (from yew tree bark), who originally studied its therapeutic activity in rodent tumours (Wani et al., 1971). One of the most common characteristics of cancer cells is that they divide rapidly and Paclitaxel acts by inhibiting this cell replication. Paclitaxel can now be made in the laboratory and is sold under the tradename Taxol.

Each half of each group in the breast cancer dataset contains three replicates at each of 4 time intervals for two experiments. Note that as the MCF-7/ADR cell-lines are chemoresistant they should not be affected by the Taxol treatment.

The breast cancer data are shown in figure 1.2 as images. Results for m/z values below 2,000 Daltons have been removed as background interference from sinapinic acid matrix peaks tend to produce low signal:noise ratios (Ball et al, 2002).

In summary there are six groups: ADC (MCF-7/ADR control), ADT (MCF-7/ADR Taxol treated), TDC (T47D control), TDT (T47D Taxol treated), MCC (MCF-7 control) and MCT (MCF-7 Taxol treated). SELDI-TOF scans were taken at periods of 24, 48, 72 and 96 hours and these are labelled as day 1, day 2, day 3 and day 4 respectively. For each group on each day there are 6 observations: replicates A,B and C for experiment 1 and replicates A,B and C for experiment 2.

The protocol for the experiment is described in detail by Mian et al. (2003), and the data were collected at Nottingham Trent University in the laboratory of Professor Robert Rees.

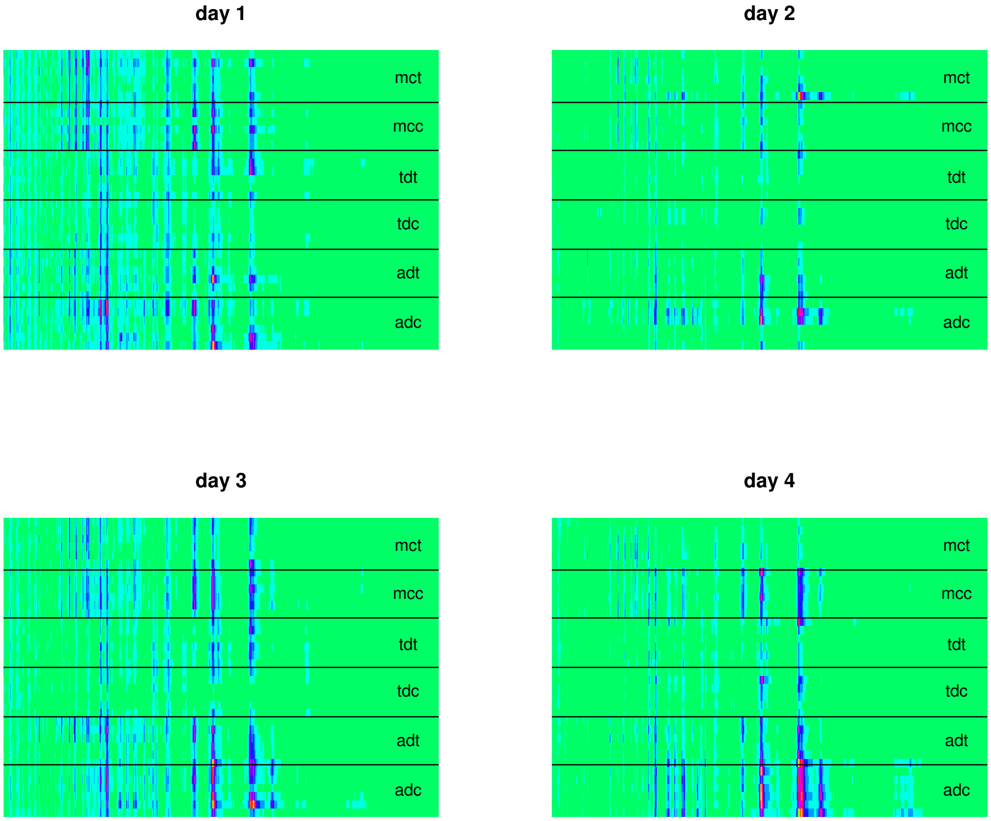

Figure 1.2: The breast cancer data. The x-axis shows m/z value starting at 2,001 Daltons and finishing at 30,000 Daltons. For each group at each m/z value there are 6 observations presented in the order (bottom to top) replicates A,B,C for experiment 1 followed by replicates A,B,C for experiment 2. Green indicates a low level of protein intensity increasing through blue and red to the highest intensity of yellow.

Melanoma is a malignant tumour of the skin and is made up of cells containing the pigment melanin. It is not very common - accounting for only about 4% of all cancer cases, however, it is one of the most serious, life-threatening forms of skin cancer accounting for 79% of skin cancer deaths (Kirkwood et al., 2003). It begins in the cells that produce the skin colouring and often appears on the skin as a new or changing mole. Melanomas are induced by exposure to high levels of UV radiation and are more common in people who have had significant sun exposure. Early stage melanoma is almost always curable, however, it is likely to spread, and once it has spread to other parts of the body the chances of a cure are much less.

The melanoma dataset consists of 205 sera - 101 of these are classed as stage I of the disease and 104 are classed as stage IV using the Tumour-lymph Node-Metastasis (TNM) system (e.g. Sobin and Wittekind, 2002). Each of these categories is given a number as per table 1.2. A lower number generally means a less serious melanoma. The value of each category is used to *stage* the melanoma by comparing with table 1.3.

The melanoma data are shown in figure 1.3 as images. As previously the m/z values below 2,000 Daltons have been removed.

The mass spectrum for a cell-line in either of the datasets consists of around 14,000 datapoints. Each datapoint comprises a relative intensity of proteins at a particular mass over charge (m/z value). The m/z value is calculated by dividing the protein mass by the number of charges induced by ionisation. We consider m/z values between 2kDa and 30kDa.

The experimental protocol is described in detail in Mian et al. (2005). The data were collected at the German Cancer Research Centre, Heidelberg in the laboratory of Dr. Dirk Schadendorf.

TNM		**What it means**
T (tumour)	T0	There is no evidence of a tumour.
	Tis	Melanoma has some of the changes that make it cancer, but it is not yet the kind that spreads into other tissues. This is also called melanoma in situ.
	T1	Melanoma is 1 millimetre or less in thickness.
	T2	Melanoma is between 1 and 2 millimetres thick.
	T3	Melanoma is between 2 and 4 millimetres thick.
	T4	Melanoma is more than 4 millimetres thick.
N (lymph nodes)	N0	Melanoma has not spread to any lymph nodes.
	N1	Melanoma has spread to one lymph node nearby.
	N2	Melanoma has spread to two or three lymph nodes nearby.
	N3	Melanoma has spread to four or more lymph nodes nearby.
M (metastasis)	M0	Melanoma has not spread to another part of the body.
	M1	Melanoma has spread to another part of the body.

Table 1.2: TNM classes for melanoma.

TNM	**Cancer Stage**
Tis, N0, M0	0
T1, N0, M0 T2, N0, M0	I
T2, N0, M0 T3, N0, M0 T4, N0, M0	II
Any T, N1, M0 Any T, N2, M0 Any T, N3, M0	III
Any T, Any N, M1	IV

Table 1.3: Stages of melanoma. Note that T2,N0,M0 can be classed as either stage I or stage II depending on its appearance under a microscope.

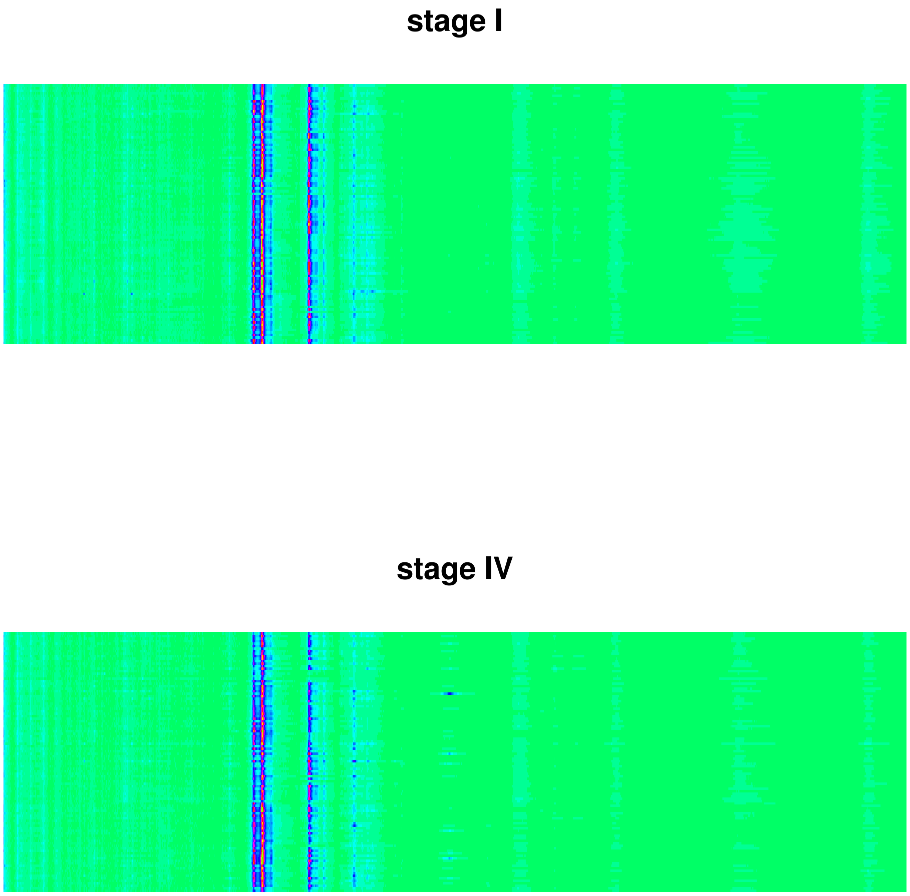

Figure 1.3: The melanoma data. The x-axis shows m/z value starting at 2,001 Daltons and finishing at 30,000 Daltons. For stage I there are 101 observations and for stage IV there are 104 observations. Green indicates a low level of protein intensity increasing through blue and red to the highest intensity of yellow.

1.4 Data Reduction and Classification

Presented in this section are the methods used in this thesis to classify observations and the two main techniques used to reduce the dimensionality of a dataset.

It is important to consider data reduction in proteomic studies. The datasets produced by time-of-flight mass spectrometry are generally high-dimensional and there are generally many more m/z values per spectrum than there are spectra.

1.4.1 Dimension Reduction Methods

Principal Components Analysis

The main goal of principal components analysis (PCA) is to explain the important variability in the data in a reduced number of dimensions. This is done by projecting the data linearly onto lower dimensional subspaces in which they show maximal variation. We describe sample principal components analysis where a sample of vectors $\mathbf{x}_1 \ldots, \mathbf{x}_n$ is available.

Let $\mathbf{u}$ be a unit vector i.e. $\mathbf{u}^T\mathbf{u} = 1$. Define $c_i = \mathbf{u}^T(\mathbf{x}_i - \bar{\mathbf{x}})$ for $i = 1 \ldots n$ where the $\mathbf{x}_i's$ are column vectors of observations, $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ is their sample mean and n is the number of observations. Now

$$\sum_{i=1}^n c_i = \sum_{i=1}^n \mathbf{u}^T(\mathbf{x}_i - \bar{\mathbf{x}}) = \mathbf{u}^T \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) = 0$$

by definition of $\bar{\mathbf{x}}$, and

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n c_i^2 &= \frac{1}{n}\sum_{i=1}^n \mathbf{u}^T(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T\mathbf{u} \\ &= \mathbf{u}^T \left[\frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T\right]\mathbf{u} \end{aligned}$$

$$= \mathbf{u}^T\bar{\mathbf{S}}\mathbf{u}. \tag{1.1}$$

So $\mathbf{u}^T\bar{\mathbf{S}}\mathbf{u}$ is the sample variance of the c_i's along $\mathbf{u}$. We would like to find the $\mathbf{u}^*$ which maximises the sample variance $\mathbf{u}^T\bar{\mathbf{S}}\mathbf{u}$ over unit vectors $\mathbf{u}$.

Since $\bar{\mathbf{S}}$, the sample covariance matrix of the observations, is symmetric, applying spectral decomposition (e.g. Mardia, Kent and Bibby, 1979) gives:

$$\bar{\mathbf{S}} = \mathbf{Q\Lambda Q^T} = \sum_{j=1}^{p} \lambda_j \mathbf{q}_j \mathbf{q}_j^T$$

where $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_p]$, $\mathbf{QQ^T} = \mathbf{Q^TQ} = \mathbf{I}_p$ and $\mathbf{\Lambda} = diag\{\lambda_1, \lambda_2, \ldots, \lambda_p\}$.

The vectors $\mathbf{q}_j$ are the eigenvectors corresponding to the eigenvalues λ_j, where $j = 1 \ldots p$. Without loss of generality assume that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Then

$$\begin{aligned}
\mathbf{u}^T\bar{\mathbf{S}}\mathbf{u} &= \mathbf{u}^T \left[\sum_{j=1}^{p} \lambda_j \mathbf{q}_j \mathbf{q}_j^T\right] \mathbf{u} \\
&= \sum_{j=1}^{p} \lambda_j \mathbf{u}^T \mathbf{q}_j \mathbf{q}_j^T \mathbf{u} \\
&= \sum_{j=1}^{p} \lambda_j (\mathbf{u}^T \mathbf{q}_j)^2 && \text{since } \mathbf{u}^T\mathbf{q}_j \text{ is a scalar} \\
&\leq \lambda_1 \sum_{j=1}^{p} (\mathbf{u}^T \mathbf{q}_j)^2 && \text{since } \lambda_1 \text{ is the largest eigenvalue} \\
&= \lambda_1 \mathbf{u}^T \mathbf{Q} \mathbf{I}_p \mathbf{Q}^T \mathbf{u} \\
&= \lambda_1 \mathbf{u}^T \mathbf{u} \\
&= \lambda_1 && \text{since } \|\mathbf{u}\| = 1 \qquad (1.2)
\end{aligned}$$

but

$$\begin{aligned}
\mathbf{q}_1^T\bar{\mathbf{S}}\mathbf{q}_1 &= \mathbf{q}_1^T\left[\sum_{j=1}^{p}\lambda_j\mathbf{q}_j\mathbf{q}_j^T\right]\mathbf{q}_1 \\
&= \sum_{j=1}^{p}\lambda_j(\mathbf{q}_1^T\mathbf{q}_j)^2 \\
&= \lambda_1. \qquad\qquad \text{since } \mathbf{q}_1^T\mathbf{q}_j = \begin{cases} 0 & j \neq 1; \\ 1 & j = 1. \end{cases} \qquad (1.3)
\end{aligned}$$

So $\mathbf{u}^T\bar{\mathbf{S}}\mathbf{u}$ is maximised over unit vectors $\mathbf{u}$ when $\mathbf{u} = \mathbf{q}_1$ i.e. the unit eigenvector corresponding to the largest eigenvalue λ_1. The largest eigenvector is unique up to sign when $\lambda_1 > \lambda_2$.

This procedure can be repeated to look for the largest sample variance of the c_i's when $\mathbf{u}$ is chosen to be orthogonal to $\mathbf{q}_1$ (i.e. when $\mathbf{u}^T\mathbf{q}_1 = 0$). Similar reasoning shows that this occurs when $\mathbf{u} = \mathbf{q}_2$, the eigenvector corresponding to the second largest eigenvalue λ_2 (e.g. Mardia, Kent and Bibby, 1979).

The values $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_p$ determine the p principal components ($p \leq n-1$). The s_{ji} are the PC scores where

$$s_{ji} = \mathbf{q}_j^T(\mathbf{x}_i - \bar{\mathbf{x}}) \qquad i = 1\ldots n, \quad j = 1\ldots p.$$

Independent Components Analysis

Independent Components Analysis (ICA) was introduced in the early 1980s in the context of neural network modelling and was developed in the 1990s with the introduction of new algorithms. ICA has been used in such applications as telecommunications, time series analysis and data mining.

ICA is a statistical technique in which observed random data are expressed as a linear combination of components that are statistically independent from each other. Such a decomposition is well defined if the independent components are non-Gaussian (Hyvärinen et al, 2001).

Assume we observe n random variables (spectra) $x_1, \ldots, x_n$ which we wish to model as linear combinations of the n random variables $s_1, \ldots, s_n$. Then

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \ldots a_{in}s_n, \text{ for all } i = 1, \ldots, n$$

where the a_{ij} $(i, j = 1, \ldots n)$ are some real coefficients. The random variables s_j are called the *independent components* (ICs). By definition the s_j are statistically mutually independent. The values of the s_j's and the a_{ij}'s are unknown and must be estimated from the observed data x_i.

This model is written in matrix notation as follows. Let $\mathbf{x}$ be the random vector consisting of the x_i and $\mathbf{s}$ be the vector consisting of the s_j. Let $\mathbf{A}$ be the matrix with elements a_{ij}. Then the model becomes

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_{i=1}^{n} \mathbf{a}_i s_i \tag{1.4}$$

where $\mathbf{a}_i$ is the i^{th} column of the matrix $\mathbf{A}$.

When using ICA we assume that the ICs are statistically independent, that the ICs have non-Gaussian distributions and that the matrix $\mathbf{A}$ must be square (same number of ICs as x_i's) and invertible. This last restriction can sometimes be relaxed (see Hyvärinen, 2001) but if it holds then we can calculate the inverse $\mathbf{A}^{-1} = \mathbf{B}$ and obtain the ICs simply by $\mathbf{s} = \mathbf{B}\mathbf{x}$

There are two problems encountered with this ICA model. Firstly we cannot determine the variances of the ICs. Since $\mathbf{A}$ and $\mathbf{s}$ are both unknown, a scalar multiplier in one can be cancelled out in the other:

$$\mathbf{x} = \sum_{i=1}^{n} \mathbf{a}_i s_i = \sum_{i=1}^{n} \frac{1}{\alpha_i} \mathbf{a}_i (s_i \alpha_i)$$

Secondly we cannot determine the order of the ICs. Again, since $\mathbf{A}$ and $\mathbf{s}$ are both unknown, the order of the terms in equation (1.4) can be rearranged and any one of them can be called the first. In practice the matrix $\mathbf{B} = \mathbf{A}^{-1}$ is estimated by calculating the vectors b_j which maximise the non-Gaussianity of the calculated s_j. Many algorithms exist which carry out ICA and in this thesis the *FastICA* algorithm is used in *R*.

1.4.2 Classification Methods

After dimension reduction has been carried out we can use the new variables to classify new observations into different disease states. This could be important to physicians when determining which treatment to give different patients. Three of the main methods for classifying data are now described.

Discriminant Analysis

In discriminant analysis there exist $G \geq 2$ populations. It is assumed that each population has a particular distribution $f_i(\mathbf{x}) \in \mathbb{R}^p$ $(i = 1, \ldots G)$. A set of data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is available for which the group membership of each observation is known. These are called the training data. Based on the assumptions and the training data, rules are constructed for assigning a new observation $\mathbf{z} \in \mathbb{R}^p$ to one of the G populations whilst minimising the probability of misclassification. The aim is to find an effective rule for discrimination based on inexpensive measurements instead of near-perfect classification using expensive measurements. An approach due to Fisher (1936) is to look for a linear discriminant function without assuming that the G populations $\Pi_1, \Pi_2, \ldots, \Pi_G$ are normally distributed.

Suppose we have training samples $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_i}$ from population $\Pi_i, i = 1 \ldots G$. Then the 'within' sum of squares covariance matrix is

$$\mathbf{W} = \sum_{i=1}^{G} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T \qquad \text{where} \qquad \bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij}$$

and the 'between' sum of squares covariance matrix is

$$\mathbf{B} = \sum_{i=1}^{G} n_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \qquad \text{where} \qquad \bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{G} \sum_{j=1}^{n_i} \mathbf{x}_{ij} \text{ and } N = \sum_{i=1}^{G} n_i$$

Fisher's criterion is to choose a unit vector $\boldsymbol{\lambda}$ to maximise

$$\frac{\boldsymbol{\lambda}^{\mathbf{T}} \mathbf{B} \boldsymbol{\lambda}}{\boldsymbol{\lambda}^{\mathbf{T}} \mathbf{W} \boldsymbol{\lambda}}. \tag{1.5}$$

Then the function $\mathcal{L}(\mathbf{z}) = \boldsymbol{\lambda}^T \mathbf{z}$ is called Fisher's linear discriminant function. To find $\boldsymbol{\lambda}$ to maximise equation (1.5) assume that $\mathbf{W}$ is positive definite and note that $\mathbf{W}$ is symmetric. So using spectral decomposition (e.g. Mardia, Kent and Bibby, 1979) gives $\mathbf{W} = \mathbf{Q} \boldsymbol{\Lambda} \mathbf{Q}^T$ and $\mathbf{W}^{1/2} = \mathbf{Q} \boldsymbol{\Lambda}^{1/2} \mathbf{Q}^T$ where $\overset{n \times n}{\boldsymbol{\Lambda}} = diag\{\mu_1, \mu_2, \ldots, \mu_n\}$ are the eigenvalues of $\mathbf{W}$ and $\overset{n \times n}{\mathbf{Q}}$ is an orthogonal matrix whose columns are the eigenvectors of $\mathbf{W}$.

Define $\boldsymbol{\gamma} = \mathbf{W}^{1/2} \boldsymbol{\lambda}$, then $\boldsymbol{\lambda} = \mathbf{W}^{-1/2} \boldsymbol{\gamma}$ where $\mathbf{W}^{-1/2} = \mathbf{Q} \boldsymbol{\Lambda}^{-1/2} \mathbf{Q}^T$. Now

$$\begin{aligned} \max_{\boldsymbol{\lambda}:\boldsymbol{\lambda}\boldsymbol{\lambda}^T=1} \left\{ \frac{\boldsymbol{\lambda}^T \mathbf{B} \boldsymbol{\lambda}}{\boldsymbol{\lambda}^T \mathbf{W} \boldsymbol{\lambda}} \right\} &= \max_{\boldsymbol{\gamma}:\boldsymbol{\gamma} \neq 0} \left\{ \frac{\boldsymbol{\gamma}^T \mathbf{W}^{-1/2} \mathbf{B} \mathbf{W}^{-1/2} \boldsymbol{\gamma}}{\boldsymbol{\gamma}^T \underbrace{\mathbf{W}^{-1/2} \mathbf{W} \mathbf{W}^{-1/2}}_{\mathbf{I}_p} \boldsymbol{\gamma}} \right\} \\ &= \max_{\boldsymbol{\gamma}:\boldsymbol{\gamma}^T\boldsymbol{\gamma}=1} \left\{ \boldsymbol{\gamma}^T \mathbf{W}^{-1/2} \mathbf{B} \mathbf{W}^{-1/2} \boldsymbol{\gamma} \right\} \end{aligned} \tag{1.6}$$

To solve this maximisation choose $\boldsymbol{\gamma}$ to be the eigenvector corresponding to the largest eigenvalue of $\mathbf{W}^{-1/2} \mathbf{B} \mathbf{W}^{-1/2}$. This is true since if $\boldsymbol{\gamma}$ is an eigenvector of $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ then by definition

$$\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}\boldsymbol{\gamma} = \rho\boldsymbol{\gamma}$$

for some constant ρ. If we premultiply by $\mathbf{W}^{-1/2}$ we get the following:

$$\begin{aligned} \mathbf{W}^{-1/2}\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}\boldsymbol{\gamma} &= \rho\mathbf{W}^{-1/2}\boldsymbol{\gamma} \\ \mathbf{W}^{-1}\mathbf{B}\left[\mathbf{W}^{-1/2}\boldsymbol{\gamma}\right] &= \rho\mathbf{W}^{-1/2}\boldsymbol{\gamma} \\ \Rightarrow \qquad \mathbf{W}^{-1}\mathbf{B}\boldsymbol{\lambda} &= \rho\boldsymbol{\lambda} \qquad (1.7) \end{aligned}$$

So the $\boldsymbol{\lambda}$ required is the unit eigenvector corresponding to the largest eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$. Fisher's linear discriminant function $\mathcal{L}(\mathbf{z})$ can now be calculated and an observation $\mathbf{z}$ will be allocated to the Π_i whose discriminant score $\mathcal{L}(\bar{\mathbf{x}}_i)$ is closest to $\mathcal{L}(\mathbf{z})$ i.e.

$$\text{allocate } \mathbf{z} \text{ to } \Pi_i \textit{ iff } |\boldsymbol{\lambda}^T\mathbf{z} - \boldsymbol{\lambda}^T\bar{\mathbf{x}}_i| = \min_{1\leq j\leq G}|\boldsymbol{\lambda}^T\mathbf{z} - \boldsymbol{\lambda}^T\bar{\mathbf{x}}_j|$$

In linear discriminant analysis, the covariance matrix of each group is assumed to be equal. In this case the decision boundaries calculated from the above equations are linear. If this assumption is not true then quadratic discriminant analysis (QDA) can be used and the decision boundaries are quadratic curves. QDA provides superior results if the group covariances are considerably different and the group sizes are large. However, QDA is more sensitive to deviations from normality and classification errors in the training set (Lachenbruch, 1982).

Support Vector Machines

Support vector machines (SVMs) were introduced by Boser, Guyon and Vapnik (1992) as a means of classifying data. They have been used in many fields including bioinformatics and image recognition. The simplest SVM is called the *maximal margin classifier* and only works for data that can be linearly separated and is therefore not valid in many real life situations. However, in 1995 a modified maximum margin idea was suggested for when the datapoints cannot be separated without error (Cortes and Vapnik, 1995). If no linear classifier exists that can correctly split the datapoints, the *soft margin* method will choose a classifier that splits the datapoints as cleanly as possible, while still maximising the distance to the nearest cleanly split datapoints.

The *margin* of a linear classifier is defined to be the width that the boundary could be increased before hitting any datapoints. In figure 1.4, for example, the classifier on the left has a smaller margin than the classifier on the right as extending the line will reach the point at (6, 9.5) more quickly.

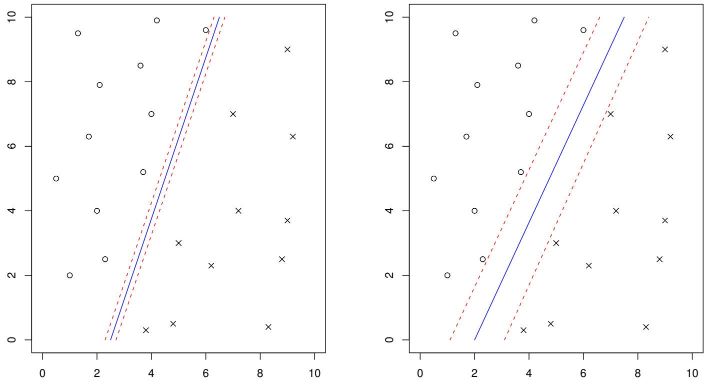

Figure 1.4: Two linear classifiers with different margins.

The classifier with the largest margin is called the *maximum margin linear classifier* and the *support vectors* are the datapoints that the maximum margin pushes up against. The use of maximum margin classifiers gives the least chance of causing a misclassification if there is a small error in the location of the boundary.

The equation of the maximum margin linear classifier in N dimensions is of the form

$$\boldsymbol{\beta}^T \mathbf{x} + \beta_0 = 0$$

and the distance from the classifying boundary to the nearest points is C. We wish to maximise the distance C whilst allowing for some datapoints to be on the wrong side of the boundary. We define the slack variables $\xi_1, \ldots, \xi_N$ as the amounts that the datapoints $\mathbf{x}_i, i = 1, \ldots, N$ are on the wrong side of the margin. We also wish to minimise the sum of these errors. The maximisation to be calculated is:

$$\max_{\beta_0, \boldsymbol{\beta}, \|\boldsymbol{\beta}\|=1} C \tag{1.8}$$

subject to the conditions $y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) \geq C(1 - \xi_i),\ i = 1, \ldots, N,\ \xi_i \geq 0,\ \sum_{i=1}^N \xi_i \leq$ constant , where the $\mathbf{x}_i$ are the datapoints and the y_i denote to which class the $\mathbf{x}_i$ belong. Note misclassifications occur when $\xi_i > 1$. We can remove the $\|\boldsymbol{\beta}\| = 1$ constraint by replacing the first condition with

$$y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) \geq C\|\boldsymbol{\beta}\|, \quad i = 1, \ldots, N$$

Since any scalar multiples of $\boldsymbol{\beta}$ and β_0 also satisfy the inequalities, we can choose to set $\|\boldsymbol{\beta}\| = 1/C$. So equation (1.8) is equivalent to

$$\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_{i=1}^N \xi_i \tag{1.9}$$

subject to $y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) \geq C(1 - \xi_i), \quad i = 1, \ldots, N,\ \xi_i \geq 0$. To solve this constrained minimisation we use the Lagrange multiplier technique (e.g. Winston, 1995). The Lagrangian is

$$\frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_{i=1}^N \xi_i - \sum_{i=1}^N \lambda_i \left[y_i \left(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0\right) - (1 - \xi_i)\right] - \sum_{i=1}^N \mu_i \xi_i \tag{1.10}$$

Differentiating equation (1.10) with respect to β_0 and setting equal to zero gives $\sum_{i=1}^N \lambda_i y_i = 0$, differentiating with respect to $\boldsymbol{\beta}$ gives $\sum_{i=1}^N \lambda_i y_i \mathbf{x}_i = \boldsymbol{\beta}$ and differentiating with respect to ξ_i gives $\lambda_i = \gamma - \mu_i, \forall i$. (Note that $\lambda_i, \mu_i, \xi_i \geq 0$). Substituting these results into 1.10 gives

$$\frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i^T \boldsymbol{\beta} - \sum_{i=1}^{N} \lambda_i y_i \beta_0 + \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \lambda_i \xi_i - \sum_{i=1}^{N} \mu_i \xi_i$$

$$\begin{aligned} = \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N} \lambda_i \lambda_k y_i y_k \mathbf{x}_i^T \mathbf{x}_k + \gamma \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N}\sum_{k=1}^{N} \lambda_i \lambda_k y_i y_k \mathbf{x}_i^T \mathbf{x}_k - \sum_{i=1}^{N} \lambda_i y_i \beta_0 \\ & \quad + \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \gamma \xi_i + \sum_{i=1}^{N} \mu_i \xi_i - \sum_{i=1}^{N} \mu_i \xi_i \end{aligned}$$

$$= \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N} \lambda_i \lambda_k y_i y_k \mathbf{x}_i^T \mathbf{x}_k - \sum_{i=1}^{N}\sum_{k=1}^{N} \lambda_i \lambda_k y_i y_k \mathbf{x}_i^T \mathbf{x}_k + \sum_{i=1}^{N} \lambda_i$$

$$= \quad \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N} \lambda_i \lambda_k y_i y_k \mathbf{x}_i^T \mathbf{x}_k \qquad (1.11)$$

Although this leaves us with another optimisation, equation (1.11) can be minimised more easily than the original problem. Firstly we differentiate and solve equation (1.11) for each λ_i. Then $\boldsymbol{\beta}$ can be determined from the equation $\sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i = \boldsymbol{\beta}$.

To determine β_0 we must use the Karush-Kuhn-Tucker condition (e.g. Winston, 1995)

$$\lambda_i[y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) - (1 - \xi_i)] = 0 \quad \forall i. \qquad (1.12)$$

There must exist at least one support vector $\mathbf{x}_i$. For all support vectors $\xi_i = 0$ as they lie on the margin. Substituting in our known values relevant to one support vector gives us the equation of the classifier. Test points can then be classified according to the side of the boundary on which they fall.

Although the classifier has been constructed in order that no training points fall within the margin, this may not necessarily hold for test points. It is hoped that a large margin separating the training data will give good test classification. When there are more than two groups into which points can be classified more than one linear classifier must be trained. In this thesis the *'one against one'* approach is used where, with k groups, each of the $k(k-1)/2$ pairwise comparisons are considered and a voting scheme determines the classification of a test point.

K-Nearest-Neighbour

The K-nearest neighbour (KNN) algorithm is a relatively simple method for classifying observed data where class memberships are known for a set of training data. For a particular observed datapoint the nearest K training points, usually using Euclidian distance, are considered and a majority vote of their K classes is taken as the class of the new datapoint. If there is a tie then it is broken at random between the tied groups. The accuracy of the KNN algorithm can be severely affected by the presence of outliers, especially when $K = 1$. The best choice of K depends upon the data. Generally, larger values of K reduce the effect of noise on the classification, but the boundaries between classes become less distinct. See Hastie et al. (2001) for further details.

1.5 MCMC for Bayesian Inference

The use of Bayesian methods in applied problems increased greatly at the end of the 20^{th} century. The availability of fast computers was combined with the development of Markov Chain Monte Carlo (MCMC) algorithms, a group of simulation methods, which allowed the study of more complex Bayesian models. The idea behind MCMC is to simulate approximate samples from the posterior distribution of interest by generating a Markov chain which has the posterior distribution as its limiting equilibrium distribution. This approach originated in the statistical physics literature (Metropolis et al., 1953) and it was then generalised by Hastings (1970). However, it was Gelfand and Smith (1990) that brought MCMC methods to the attention of the general statistical community, and since then the use of Bayesian methods for applied statistical modelling has increased rapidly.

Gilks et al. (1996) gives an overview of advances in MCMC related methodology until 1995. MCMC software is also being produced and made freely available to analyse a wide range of statistical models. An example is BUGS - *B*ayesian inference *U*sing *G*ibbs *S*ampling (see Spiegelhalter et al., 1996). We will use MCMC in chapter 3.

1.5.1 Bayesian Inference

The fundamentals of Bayesian theory are reviewed in this section in an introductory manner. For a more detailed approach see Bernardo and Smith (1994).

Bayes' Theorem

In classical inference the data, which are assumed to depend on a vector of parameters, $\boldsymbol{\theta}$, are thought of as random with $\boldsymbol{\theta}$ fixed (but unknown). In Bayesian inference the thinking is opposite - the data are regarded as fixed (at what has been observed) and the parameter vector $\boldsymbol{\theta}$ is treated as unknown.

In the Bayesian approach in addition to specifying the model for the observed data $\mathbf{y} = (y_1, \ldots, y_n)$ given the vector of unknown parameters $\boldsymbol{\theta}$, in the form of the likelihood function $\pi(\mathbf{y}|\boldsymbol{\theta})$, we also define the *prior* distribution $\pi(\boldsymbol{\theta})$. The prior should contain all knowledge we have about the unknown parameter before analysis starts. Inference concerning $\boldsymbol{\theta}$ is then based on its *posterior* distribution, given by

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{1.13}$$

This formula is referred to as *Bayes' Theorem.* The integral in the denominator is a normalising constant to ensure the distribution is a valid probability distribution and it's calculation has traditionally been a computational obstacle. The main difficulty is that the calculation involves a many-dimensional integration and the resulting distribution cannot always be written down in closed form. However, it is possible to avoid its calculation using MCMC methods. Equation (1.13) can be thought of as *"The posterior is proportional to the likelihood times the prior"*.

Prior Distributions

Presented here are the two most popular approaches for choosing a prior distribution.

Informative priors

An informative prior for a parameter θ is a prior used when some information is known about the parameter before any data is obtained. For example, assume we were interested in estimating the average weight of newborn female babies. Then before we actually collect any observations of the weight of newborn babies, we find on a website that the average weight of a newborn is 3.4 kg. A prior is then chosen to incorporate this information - we choose a normal prior with mean 3.4 and variance σ^2. The value of σ^2 is still to be chosen and will incorporate the strength of our belief in the mean value of 3.4. The lower the value of σ^2, the stronger our belief in the mean. If we believe female babies to weigh less than male babies we can also include this belief by reducing our mean value.

Non-Informative or Diffuse Priors

In many situations no prior information concerning θ is available, or inference based solely on the data is desirable. Typically in this case we wish to define a prior distribution $\pi(\theta)$ that contains no information whatsoever about the parameter θ in the sense that it does not favour one particular value of θ over another. Such a distribution is called a *noninformative prior* for θ and it can be argued that the information about θ contained in the posterior comes only from the data. In classical inference prior distributions are not used in fitting models and so 'noninformative' priors are often used in Bayesian inference to compare with classical results.

In the case where the parameter space is $\mathbf{\Theta} = \{\theta_1, \ldots, \theta_n\}$ i.e. discrete and finite, then the distribution

$$\pi(\theta_i) = \frac{1}{n}, \quad i = 1, \ldots, n$$

places the same prior probability of $1/n$ on any of the n candidate θ values. Similarly, in the case of a bounded continuous parameter space, say
$\mathbf{\Theta} = [a, b], -\infty < a < b < \infty$, then the uniform distribution

$$\pi(\theta) = \frac{1}{b-a}, \quad a < \theta < b$$

is noninformative. A normal distribution with large variance may also be used as a noninformative prior. As the variance of a normal distribution is increased, the distribution becomes 'flatter' around the mean (see figure 1.5). This explains the alternative names for noninformative priors of 'flat' or 'diffuse' priors.

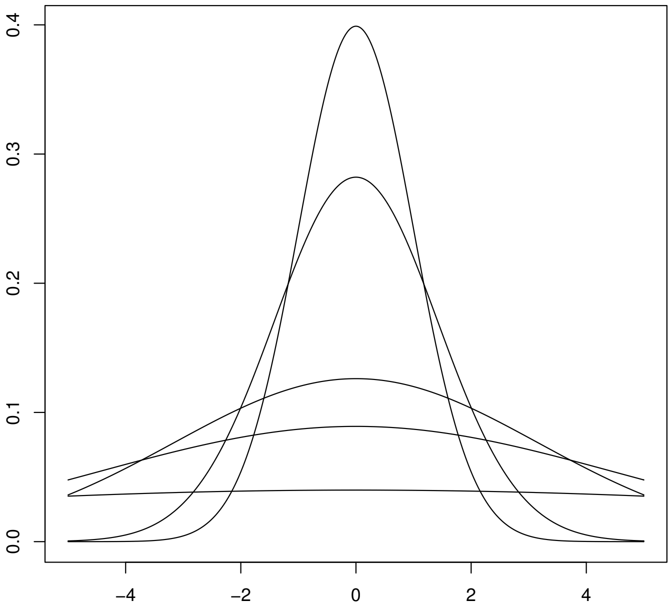

Figure 1.5: Normal distributions with mean 0 and variances 1, 5, 10, 20 and 100 respectively.

For unbounded intervals the definition of a noninformative distribution is not straightforward. In the case that $\boldsymbol{\Theta} = (-\infty, \infty)$ a distribution such as $\pi(\theta) = c$ is clearly improper since $\int \pi(\theta) d\theta = \infty$. However, Bayesian inference is still possible in the case where $\int \pi(\mathbf{y}|\theta) d\theta = D < \infty$. Then

$$\pi(\theta|\mathbf{y}) = \frac{\pi(\mathbf{y}|\theta)c}{\int \pi(\mathbf{y}|\theta) c d\theta} = \frac{\pi(\mathbf{y}|\theta)}{D} \ .$$

It should be noted that there is not a ‘universal’ noninformative prior. It is possible in some cases for a constant prior to actually be informative under a different parameterisation. One method used to overcome this problem is the use of *Jeffreys prior*. Jeffreys prior (1946) takes the form $\pi(\theta) \propto I(\theta)^{1/2}$, where

$$I(\theta) = \mathbb{E}\left[\left[\frac{\partial}{\partial\theta} \log \pi(y|\theta)\right]^2\right]$$

is the Fisher information.

When choosing a prior from a parametric family it can be possible to select a distribution which is *conjugate* to the likelihood, that is one that leads to a posterior belonging to the same family as the prior. The use of MCMC does not require conjugate priors but they can be computationally convenient.

1.5.2 Markov Chain Monte Carlo

We will now consider a collection of algorithms that greatly facilitate the implementation of Bayesian modelling known as Markov Chain Monte Carlo (MCMC) algorithms.

Suppose a sequence of random variables $\{X_0, X_1, X_2, \ldots\}$ is generated such that at each time $t \geq 0$, the next state X_{t+1} is sampled from a distribution $P(X_{t+1}|X_t)$ which depends only on the current state of the chain X_t. So given X_t, the next state X_{t+1} does not depend on the remainder of the history of the chain $\{X_0, X_1, \ldots, X_{t-1}\}$. This sequence is called a Markov chain.

The main idea behind MCMC is to generate a Markov chain which has as its limiting equilibrium distribution the posterior distribution of interest. For MCMC to work the chain must be aperiodic, irreducible and reversible (Gilks et al., 1996). MCMC first appeared in Metropolis et al. (1953) although the computational power was not available then to carry out many of the procedures used today. The original mechanism was generalised by Hastings (1970) in the *Metropolis-Hastings* algorithm, which is now described.

The Metropolis-Hastings Algorithm

The objective of the Metropolis-Hastings (MH) algorithm is to generate approximate samples from a posterior density $\pi(\theta|y)$ known up to a normalising constant. For this algorithm, at each time t, the next state $\theta^{(t+1)}$ is chosen by first sampling a *candidate* point ϕ from a *proposal distribution* $q(.|\theta^{(t)})$. Note that the proposal distribution may depend on the current state $\theta^{(t)}$. The candidate point is then *accepted* with probability $\alpha(\theta^{(t)}, \phi)$ where

$$\alpha(\theta^{(t)}, \phi) = min\left\{1, \frac{\pi(\phi|y)q(\theta^{(t)}|\phi)}{\pi(\theta^{(t)}|y)q(\phi|\theta^{(t)})}\right\}$$

If a candidate point is accepted, the next state becomes $\theta^{(t+1)} = \phi$ and if the candidate point is rejected the chain remains in the same place i.e. $\theta^{(t+1)} = \theta^{(t)}$. So the algorithm generates a Markov chain $(\theta^{(t)})$ through the following steps:

1. Start with an arbitrary initial set of parameter values $\theta^{(0)}$.
2. Update from $\theta^{(t)}$ to $\theta^{(t+1)}$ $(t = 0, 1, \ldots)$ by
   (a) Generate $\phi \sim q(.|\theta^{(t)})$
   (b) Evaluate $\alpha(\theta^{(t)}, \phi) = min\left\{1, \frac{\pi(\phi|y)q(\theta^{(t)}|\phi)}{\pi(\theta^{(t)}|y)q(\phi|\theta^{(t)})}\right\}$
   (c) Sample a point U from a Uniform$(0, 1)$ distribution.
   (d) Set
   $$\theta^{(t+1)} = \begin{cases} \phi & \text{If } U \leq \alpha(\theta^{(t)}, \phi); \\ \theta^{(t)} & \text{otherwise.} \end{cases}$$

Metropolis-Hastings updates can be carried out in different ways. Firstly all the parameters can be updated at the same time so the candidate point ϕ would become a vector of parameters. Either all the parameters are accepted or they are all rejected. Secondly, they can be updated one parameter at a time and each iteration t would comprise of n updates (where there are n parameters). Lastly a combination could be used where the parameters are split into blocks and each block of parameters is updated at the same time.

The Gibbs Sampler

The Gibbs sampling approach is a special case of the Metropolis-Hastings algorithm. The method derives its name from Gibbs random fields, where it was used for the first time by Geman and Geman (1984).

Suppose we have a vector $\boldsymbol{\theta}$ consisting of n parameters $\theta_1, \ldots, \theta_n$. Consider the conditional density of θ_i given the data y and all the other elements of $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}_{(-i)}$ be the vector $\boldsymbol{\theta}$ with element θ_i removed. Then the distributions $\pi_i(\theta_i | \boldsymbol{\theta}_{(-i)}, y)$ for $i = 1, \ldots, n$ are called the *full conditional distributions* of $\pi(\boldsymbol{\theta}|y)$.

The idea of Gibbs sampling is to sample from the joint posterior distribution $\pi(\theta_1, \theta_2, \ldots, \theta_n|y)$ using the full conditional distributions. The parameters θ_i are updated by sampling from each of the full conditionals in turn, cycling round the parameters in each iteration. Start with an initial parameter vector $\boldsymbol{\theta}^{(0)}$ and then generate $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ as follows. Given the current state of the chain $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, ..., \theta_n^{(t)})$

$$
\begin{array}{l}
\text{Generate } \theta_1^{(t+1)} \text{ from } \pi(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}, \theta_4^{(t)}, ..., \theta_n^{(t)}, y) \\
\text{Generate } \theta_2^{(t+1)} \text{ from } \pi(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \theta_4^{(t)}, ..., \theta_n^{(t)}, y) \\
\text{Generate } \theta_3^{(t+1)} \text{ from } \pi(\theta_3|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \theta_4^{(t)}, ..., \theta_n^{(t)}, y) \\
\vdots \\
\text{Generate } \theta_n^{(t+1)} \text{ from } \pi(\theta_n|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \theta_3^{(t+1)}, ..., \theta_{n-1}^{(t+1)}, y)
\end{array}
$$

We now have the next state in the chain $\boldsymbol{\theta}^{(t+1)}$. The distribution of $\boldsymbol{\theta}^{(t)}$ tends to $\pi(\boldsymbol{\theta}|y)$.

The Gibbs sampler is a special case of the Metropolis-Hastings sampler and has acceptance probability 1. Assume again that we have parameters $\theta_1, \ldots, \theta_n$ and we wish to estimate the posterior density $\pi(.)$. The proposal distribution in a Gibbs sampling update is the full conditional distribution of the parameter i.e. $q(\theta_i|\boldsymbol{\theta}_{-i}) = \pi_i(\theta_i|\boldsymbol{\theta}_{-i})$. Let $\boldsymbol{\theta} = (\theta_1^t, \ldots, \theta_n^t)$ be the current state of the chain and let $\boldsymbol{\theta}' = (\theta_1^{t+1}, \ldots, \theta_n^t)$ be the proposed state of the chain when updating parameter θ_1. Note that the full conditional distribution of the parameter θ_1 can be expressed through the following equation:

$$\pi(\theta_1, \theta_2 \ldots, \theta_n) = \pi_1(\theta_1|\theta_2 \ldots, \theta_n)\pi(\theta_2 \ldots, \theta_n)$$

The acceptance probability for a Metropolis-Hastings update as given in the section above is:

$$\begin{aligned}
\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') &= \min\left\{1, \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\} \\
&= \min\left\{1, \frac{\pi(\theta_1^{t+1}, \theta_2^t \ldots, \theta_n^t)\pi_1(\theta_1^t|\theta_2^t \ldots, \theta_n^t)}{\pi(\theta_1^t, \theta_2^t \ldots, \theta_n^t)\pi_1(\theta_1^{t+1}|\theta_2^t \ldots, \theta_n^t)}\right\} \\
&= \min\left\{1, \frac{\pi(\theta_1^{t+1}, \theta_2^t \ldots, \theta_n^t)\pi(\theta_1^t, \theta_2^t \ldots, \theta_n^t)\pi(\theta_2^t \ldots, \theta_n^t)}{\pi(\theta_1^t, \theta_2^t \ldots, \theta_n^t)\pi(\theta_1^{t+1}, \theta_2^t \ldots, \theta_n^t)\pi(\theta_2^t \ldots, \theta_n^t)}\right\} \\
&= 1.
\end{aligned}$$

It is possible to use a mixture of updating methods for the parameter values within the same iteration of the MCMC - Gibbs sampling for some of the parameters and Metropolis-Hastings for the remainder.

Proposal Distributions

Metropolis-Hastings samplers require a proposal distribution which is used to simulate the next parameter value. A proposal distribution is usually dependent on the immediately previous value of the parameter, but independent of other previous values to ensure the Markov property holds.

A popular choice of proposal distribution is the *random walk* proposal. A common example is the normal distribution centered at the current parameter value. The proposal variance is arbitrary in this distribution and the value assigned to it affects the simulation. If the variance is too small then lots of small updates will be made to the parameter but it may take a long time to reach all areas of the sample space. If the variance is too large this will result in many proposals being rejected and the chain will not move very often at all. In both cases this means good estimates of the parameter will take a long time to achieve. See figure 1.6 for examples of plots of parameter histories when the proposal distributions have either too small or too large variance. We wish to strike a happy medium between these two cases where the proposed points are accepted around 40-60% of the time.

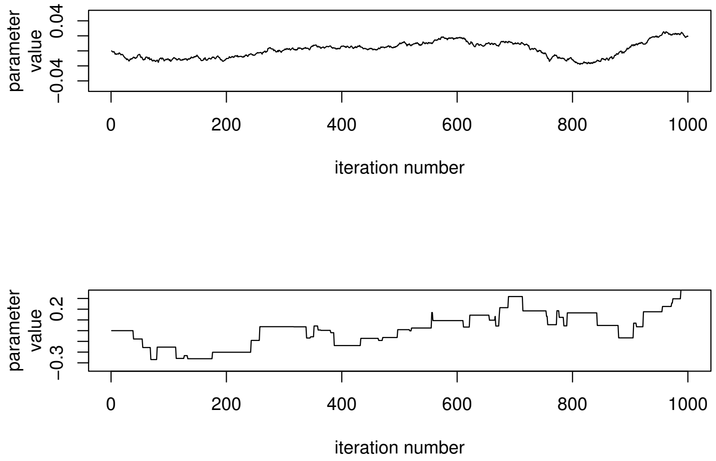

Figure 1.6: TOP: plot of a parameter history where the variance of the proposal distribution is too small. Note that nearly all proposals are accepted. BOTTOM: plot of a parameter history where the variance of the proposal distribution is too large. Note that only a few proposals are accepted.

Symmetric proposal distributions simplify the computation in parameter updates using the Metropolis-Hastings algorithm. The ratio to be calculated in a Metropolis-Hastings update is

$$\alpha(\theta^{(t)}, \phi) = \min\left\{1, \frac{\pi(\phi|y)q(\theta^{(t)}|\phi)}{\pi(\theta^{(t)}|y)q(\phi|\theta^{(t)})}\right\}$$

but if, for example, the proposal distribution is Gaussian centred around the current estimate, the ratio of proposal distributions is

$$\frac{q(\theta^{(t)}|\phi)}{q(\phi|\theta^{(t)})} = \frac{N(\phi,\sigma^2)}{N(\theta^{(t)},\sigma^2)} = \frac{\dfrac{1}{\sqrt{2\pi\sigma^2}}e^{-\dfrac{1}{2\sigma^2}(\theta^{(t)}-\phi)^2}}{\dfrac{1}{\sqrt{2\pi\sigma^2}}e^{-\dfrac{1}{2\sigma^2}(\phi-\theta^{(t)})^2}} = 1$$

thus simplifying the Metropolis-Hastings update ratio to

$$\alpha(\theta^{(t)},\phi) = \min\left\{1, \frac{\pi(\phi|y)}{\pi(\theta^{(t)}|y)}\right\}$$

Hence, the chain will remain in states with higher posterior probability more often while states with lower posterior probability are visited less often. This special case of the Metropolis-Hastings algorithm was the original case proposed by Metropolis et al. (1953).

Convergence, Burn In and Thinning

MCMC has enabled the application of Bayesian methods to many situations in different branches of study. However, to ensure that the results from MCMC algorithms are reliable two important issues need to be taken into consideration - *burn in* and *thinning*.

For a Markov chain whose distribution $(\theta^{(t)})$ converges as $t \rightarrow \infty$ there is a sufficiently large t, such that the resulting $(\theta^{(t)})$ is an approximate observation from the posterior distribution $\pi(\theta|y)$. However, the speed at which this happens, i.e. the *rate of convergence* of the chain, can vary greatly. The *burn in* period of a chain consists of all the iterations up to iteration t and these iterations are discarded from the analysis. From then on, the parameter values obtained are sampled approximately from the posterior distribution of interest provided the conditions of irreducibility, aperiodicity and reversibility are satisfied (see for example, Gilks et al., 1996).

The main problem comes in determining how large *'t'* has to be. This is generally monitored by the use of *trace plots.* Trace plots are plots of the history of the parameter values over many iterations and an example of a trace plot where convergence has been reached is shown in figure 1.7.

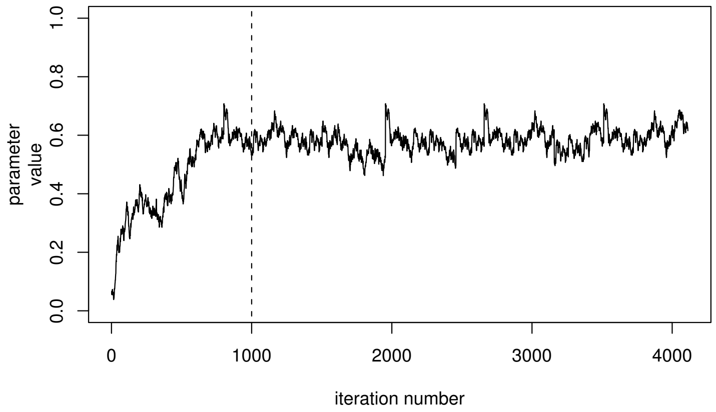

Figure 1.7: A trace plot of a model parameter. The dashed line shows roughly where the burn in period ends. Iterations to the left of the line are discarded as *burn in*, iterations to the right are used for analysis.

So after the burn in period we obtain, at each iteration, a sample from the posterior distribution $\pi(\theta|y)$. However, the samples obtained are not usually independent observations. To reduce the auto-correlation between observations we can *thin* the chain. This means that only one observation is kept for analysis every k iterations where the value of k can be chosen by the experimenter. Thus approximate independent sampling from $\pi(\theta|y)$ can be achieved. Also the amount of data created is reduced which aids the analysis.

1.6 Multilevel Modelling

There are many real-life examples of multilevel structures. The most popular example is that of an educational system with pupils within classes within schools. We say that pupils are *nested* within classes which are *nested* within schools. Goldstein and Spiegelhalter (1996) use multilevel modelling in this educational setting.

In the proteomic setting, the *m/z* values are called *level 1 units*. This is the lowest level of classification. The *level 2 units* are the spectra to which the *m/z* values belong. The data structure for a general dataset can theoretically contain any number of levels.

When data conform to such a multilevel structure it is important to take account of this in the analysis otherwise the results obtained could be inaccurate. For example, intensities at *m/z* values close to one another within one spectrum are more likely to be similar than those from another spectrum. This means they provide less information than if they were independent observations. It is important to know how this structure in the data will affect the analysis. Before the development of multilevel modelling, the problems of ignoring hierarchical structures were understood but they were difficult to solve because general tools were not available. In order to see why multilevel modelling is important we firstly review linear modelling.

1.6.1 Linear Modelling

Linear modelling is concerned with explaining the relationship between a single response variable Y and one or more predictor variables $X_1, X_2, \ldots, X_p$. A linear model is written as

$$\begin{aligned} y_i &= \mathbf{x}_i^T \boldsymbol{\beta} + e_i \\ &= x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots + x_{pi}\beta_p + e_i \ , \ (i = 1, \ldots, n) \qquad (1.14) \end{aligned}$$

In equation (1.14), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of unknown, fixed parameters. The data consists of n observations each comprising a response y_i and p predictors $\mathbf{x}_i$. It is assumed that there exists only one random error term e_i for each observation and that these errors are *iid* $N(0, \sigma^2)$ where σ^2 is unknown.

The simplest linear model is the null model and is written simply as

$$y_i = \beta_0 + e_i \ , \ e_i \sim N(0, \sigma^2).$$

In this model we are only finding the mean (β_0) and the variance (σ^2) of the sample. To improve the fit of the model we can include a continuous predictor variable. This creates a linear regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + e_i \ , \ e_i \sim N(0, \sigma^2).$$

This fits a regression line relating Y to X. Other variables can also be added so that the response variable Y is explained by more than one predictor variable.

Categorical predictors, where the values identify group membership, can be included in a linear model by considering an *Analysis of Variance* (ANOVA) model. Here the categorical predictor divides the data into J groups. The ANOVA model is written as

$$y_i = \beta_0 + \sum_{k=1}^{J} d_{ki}\alpha_k + e_i \ , \ e_i \sim N(0, \sigma^2)$$

where $d_{ki} = 1$ if $k = i$, and 0 otherwise. To make sure the model is identifiable a constraint is placed on the α's, e.g. $\alpha_1 = 0$. This model can be extended to the *Analysis of Covariance* (ANCOVA) model by additionally including a continuous predictor,

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \sum_{k=1}^{J} d_{kij}\alpha_k + e_{ij} \ , \ e_{ij} \sim N(0, \sigma^2) \ , \ \alpha_1 = 0$$

with d_{kij} defined as in the ANOVA model.

In all of the above examples of linear models the general form is

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + e_{ij}$$

and the only difference is in the definition of $\mathbf{x}_{ij}$ and $\boldsymbol{\beta}$. For the ANCOVA model $\mathbf{x}_{ij}^T = (1, x_{1ij}, d_{2ij}, \ldots, d_{Jij})$ and $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \alpha_2, \ldots, \alpha_J)$. Note that d_{1ij} and α_1 are omitted since $\alpha_1 = 0$.

To estimate the values of the parameters in $\boldsymbol{\beta}$ we can use *Least Squares Estimation* to obtain *Maximum Likelihood Estimates*. Writing the model in matrix form gives the following:

$$\mathbf{y} \;=\; \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \text{ , } E(\mathbf{e}) = 0 \text{ , } E(\mathbf{e}^T\mathbf{e}) = \sigma^2 \tag{1.15}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$, $\mathbf{X} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_n^T)^T$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$, $\mathbf{e} = (e_1, e_2, \ldots, e_n)^T$ and $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

The least squares method estimates the parameters by minimising the sum of the squared residuals $\sum_i (y_i - (\mathbf{X}\boldsymbol{\beta})_i)^2$ and it is easy to show that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{1.16}$$

assuming $\mathbf{X}$ has full rank. Thus estimates can be obtained for the parameters in a linear model simply by multiplying the correct matrices.

1.6.2 How Multilevel Modelling Differs from Linear Modelling

In the proteomic setting previously mentioned there exist *m/z* values within spectra. The modelling problem is how to model the intensities at each *m/z* value. Let y_{ij} be the intensity for m/z value i in spectrum j. A possible model is

$$y_{ij} = \beta_0 + u_j + e_{ij} \ , \ u_j \sim N(0, \sigma_u^2) \ , \ e_{ij} \sim N(0, \sigma_e^2) \qquad (1.17)$$

In this model β_0 is the estimated average for the group, u_j are the spectrum effects (which have variance σ_u^2) and e_{ij} are the m/z value residuals (which have variance σ_e^2). We assume that the u_j and the e_{ij} are independent and identically distributed (*iid*) and also that they are independent from each other.

In the ANOVA model the equivalent of the spectrum effects were the α_j. These parameters were considered to be fixed effects in the model and a constraint was required to fully identify the model. Conversely, in the multilevel model in equation (1.17) the spectrum effects are treated as random effects that come from a normal distribution. This model is called a *variance components model* as the total variance in the response $\mathbf{y}$ is split into two parts - a variance between spectra σ_u^2 and a variance between m/z values within spectra σ_e^2.

The next type of multilevel model is the *random intercept* model and is related to the ANCOVA linear model described in the previous section. The model is

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij} \qquad (1.18)$$

where $u_j \sim N(0, \sigma_u^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. This will produce a different regression line for each of the spectra, although each will have the same slope as the model assumes that the influence of the predictor variable x is the same for each spectrum. To remove this assumption we can consider a *random slopes* model. This is similar to a separate regression for each spectrum in linear modelling, however, the results will not be exactly the same due to the random effects assumption. In the random slopes model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{0j} + u_{1j} x_{ij} + e_{ij} \qquad (1.19)$$

where $\mathbf{u}_j = (u_{0j}, u_{1j}) \sim MVN(\mathbf{0}, \boldsymbol{\Omega}_u)$ and $e_{ij} \sim N(0, \sigma_e^2)$. The multivariate normal (MVN) distribution with a mean vector $\boldsymbol{\mu}$ of dimension $1 \times N$, and positive-definite, real, $N \times N$ covariance matrix $\boldsymbol{\Sigma}$ has probability density function

$$f_X(x_1, \ldots, x_N) = \frac{1}{(2\pi)^{N/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(x - \boldsymbol{\mu})\right).$$

We now consider a general way of expressing all multilevel models with 2 levels.

1.6.3 The General Two-Level Model

All of the multilevel models in section 1.6.2 have had two levels - *m/z* value and spectrum. The general two-level model is written thus:

$$y_{ij} = \mathbf{X}_{ij}^T \boldsymbol{\beta} + \mathbf{Z}_{ij}^T \mathbf{u}_j + e_{ij} \tag{1.20}$$

with $\mathbf{u}_j \sim MVN(0, \boldsymbol{\Omega}_u)$ and $e_{ij} \sim N(0, \sigma_e^2)$. This can be written in matrix form as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$. Alternatively it can be written in multivariate normal formulation as

$$\mathbf{y} \sim MVN(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$$

where $\mathbf{V} = \mathbf{Z}\boldsymbol{\Omega}_u\mathbf{Z}^T + \sigma_e^2\mathbf{I}$ is a variance/covariance matrix of dimension $N \times N$ (in our case N is the number of spectra multiplied by the number of *m/z* values).

In a similar way to how the parameter estimates were obtained for linear models using least squares estimation, *Generalised Least Squares* (GLS, Aitken, 1935) can be used in the multivariate case. We wish to obtain estimates of $\boldsymbol{\beta}$ in the following equation of which the above multivariate normal model is a special case:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \text{ , } E(\mathbf{e}) = \mathbf{0} \text{ , } E(\mathbf{e}\mathbf{e}^T) = \mathbf{V}.$$

The likelihood for the general two level model is

$$L(\boldsymbol{\beta}, \mathbf{V}; \mathbf{y}) \propto (2\pi)^{-N/2}|\mathbf{V}|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\}$$

and thus the loglikelihood is

$$\begin{aligned} l(\boldsymbol{\beta}, \mathbf{V}; \mathbf{y}) &\propto -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= -\frac{1}{2}\Big(\log|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\Big) \end{aligned}$$

The deviance of the model is defined as $D = -2 \times l(\boldsymbol{\beta}, \mathbf{V}; \mathbf{y})$ and is used to compare models. The model with the lower deviance is considered a better model.

It can be determined using the likelihood and GLS (e.g. Mardia, Kent and Bibby, 1979) that the solution to the maximising problem is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{1.21}$$

and the results in equation (1.16) can be obtained on substitution of $\mathbf{V} = \sigma^2\mathbf{I}$.

The matrix $\mathbf{V}$ is the covariance matrix for all observations. It is block diagonal (with each block representing one level 2 unit) and its elements depend on the model being used. For the general two level model in equation (1.20) the elements of $\mathbf{V}$ are:

- $Cov(y_{ij}, y_{ij'}) = 0$, $(i \neq i', j \neq j')$.
- $Cov(y_{ij}, y_{i'j'}) = \mathbf{Z}_{ij}^T\boldsymbol{\Omega}_u\mathbf{Z}_{i'j}$, $(i \neq i', j = j')$
- $Cov(y_{ij}, y_{i'j'}) = \mathbf{Z}_{ij}^T\boldsymbol{\Omega}_u\mathbf{Z}_{ij} + \sigma_e^2$, $(i = i', j = j')$

where i indexes the level 1 unit and j indexes the level 2 unit.

The use of block diagonal matrices is useful when using GLS to estimate parameters as it involves inverting matrices. If $\mathbf{V}$ were not block diagonal then inverting it would become difficult as its dimension increased. The block diagonal structure means that each of the blocks can be inverted separately and then combined.

1.6.4 The IGLS Algorithm

In the multilevel setting described in section 1.6.3, estimates need to be found for the fixed effects parameters $\boldsymbol{\beta}$ and the variance parameters $\boldsymbol{\Omega}_u$ and σ_e^2. If the two variances $\boldsymbol{\Omega}_u$ and σ_e^2 were known then the covariance matrix $\mathbf{V}$, containing the variances and covariances of the random terms over all levels of the data, can be calculated using GLS to be

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{1.22}$$

Since it is unlikely that these variances will be known then they will need to be estimated from the data using an iterative procedure. Iterative Generalised Least Squares (IGLS) is an iterative algorithm developed by Goldstein (1986) and is based on generalised least squares (GLS) estimation. The algorithm will now be described.

Firstly note that, if A is an $(m \times n)$ matrix, with columns, $A_1, A_2, ..., A_n$, each vectors of length m, then the vector of length mn obtained by stacking the columns on top of one another is denoted,

$$\text{vec}(A) = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_n \end{bmatrix}.$$

Secondly, if A is a matrix of dimension $(m \times n)$, with individual entries, a_{ij}, $i = 1, ..., m$, $j = 1, ..., n$; and B is a $(p \times q)$ matrix then the Kronecker product of A and B is defined as,

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{bmatrix},$$

which is an $(mp \times nq)$ matrix.

The first step in the IGLS algorithm is to assume that you have a simple linear model and obtain an initial estimate for $\boldsymbol{\beta}$:

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

Next, define the vector of residuals. For the random intercept model described in section 1.6.2 the residuals are $\tilde{y}_{ij} = y_{ij} - \hat{\beta}_0 - x_{ij}\hat{\beta}_1$. We can then combine these residuals into a vector $\tilde{\mathbf{Y}} = \{\tilde{y}_{ij}\}$ from which we can define $\mathbf{Y}^* = vec(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T)$. The matrix $\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T$ has expected value $\mathbf{V}$ and so $\tilde{\mathbf{Y}}$ allows us to model the variances.

The elements of $\mathbf{V}$ for the random intercept model are $\sigma_u^2 + \sigma_e^2$ for elements on the diagonal, σ_u^2 for other elements within the blocks on the diagonal and zero elsewhere. This gives the structure of $\mathbf{Y}^*$ as

$$\mathbf{Y}^* = \begin{pmatrix} \tilde{y}_{11}^2 \\ \tilde{y}_{21}^2\tilde{y}_{11}^2 \\ \vdots \\ \tilde{y}_{kn}^2 \end{pmatrix} = \begin{pmatrix} \sigma_u^2 + \sigma_e^2 \\ \sigma_u^2 \\ \vdots \\ \sigma_u^2 + \sigma_e^2 \end{pmatrix} + E = \sigma_u^2 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \sigma_e^2 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 1 \end{pmatrix} + E = \mathbf{Z}^*\boldsymbol{\theta} + E \tag{1.23}$$

for n level 2 units each with k level 1 units and where $\boldsymbol{\theta} = [\, \sigma_u^2 \;\; \sigma_e^2 \,]$.

To update the estimates obtained earlier from assuming a simple linear model we then iterate between

$$\hat{\boldsymbol{\theta}} = (\mathbf{Z}^{*^T}\mathbf{V}^{*^{-1}}\mathbf{Z}^{*})^{-1}\mathbf{Z}^{*^T}\mathbf{V}^{*^{-1}}\mathbf{Y}^{*}$$

and

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{T}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{V}^{-1}\mathbf{y}$$

and where $V^* = V \bigotimes V$ to get better estimates. Convergence occurs when successive iterations give estimates within a certain tolerance value of each other. Since IGLS is an extension of GLS estimation the assumption of normality is not required. If the assumption holds, however, then MLE estimates for the parameters are obtained.

1.6.5 Hypothesis Testing

The IGLS algorithm described in the last section gives estimates of both the fixed and variance parameters in a multilevel model along with their standard errors. To identify when particular parameters are important in the model we can carry out hypothesis tests to show statistical significance.

In hypothesis testing a question of interest is simplified into two mutually exclusive hypotheses between which there is a choice; the null hypothesis, H_0, and the alternative hypothesis, H_1. The null hypothesis normally represents no difference, for example, in the melanoma dataset a possible null hypothesis is that at an m/z value of 8,000 Daltons there is no difference in the intensities between stage I and stage IV. The alternative hypothesis is normally a statement of what the test is set up to establish, for example, that at an m/z value of 8,000 Daltons the relative intensity in stage I is lower than in stage IV. The appropriate test statistic is calculated and compared with the critical value (at a particular significance level) to determine which hypothesis is accepted.

For the fixed effects in a multilevel model we are normally interested in whether the parameter estimate is significantly different from zero. We have obtained the parameter estimate $\hat{\beta}_1$ and its estimated standard error $\hat{SE}(\hat{\beta}_1)$ and thus we could carry out a t-test (Student, 1908). However, since the size of the datasets normally considered for multilevel modelling are so large we can obtain an approximate result using a Z-test. In this case the value of the statistic Z is:

$$Z = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)}$$

which is distributed as N(0,1) under $H_0 : \beta_1 = 0$. Thus at the 5% level we reject the null hypothesis $H_0 : \beta_1 = 0$ in favour of the alternative hypothesis $H_A : \beta_1 \neq 0$ when $|Z| > 1.96$

It is also important to consider which model is the best for the data available. The best model is easy to identify when the models are *nested*. Two models, A and B, are said to be nested if the parameters in model A are a proper subset of the parameters of model B. In this case the null hypothesis is $H_0 : \theta = 0$ for all θ parameters in the complex model which are not in the simpler model. We calculate the deviance of each model, $Dev(A)$ and $Dev(B)$, and then the approximate test statistic under the null hypothesis is $D = (Dev(A) - Dev(B)) \sim \chi^2_p$ where p is the number of extra parameters in the more complex model (for large samples).

Another method which is used to compare models is the Akaike Information Criterion (AIC, Akaike, 1974). This method more directly takes into account the number of extra parameters in a model. The AIC statistic λ_A for a model A is calculated by $\lambda_A = Dev(A) + 2p$ where p is the numbers of parameters in the model. A lower value of λ_A indicates a more preferable model. The AIC criterion does not require the use of nested models.

A similar criterion to AIC is the Bayesian Information Criterion (BIC, Schwartz, 1978). The BIC statistic λ_B for a model B is calculated by $\lambda_B =$

$Dev(B)$+$p$log(n) where p is the number of parameters in the model and n is the number of datapoints. For the datasets considered in this thesis the values of n will be 2,009,088 and 2,859,955 for the breast cancer and melanoma datasets respectively and thus the value of log(n) for both datasets is around 15 as compared with 2 in the AIC statistic. This means that the introduction of more parameters is punished more harshly when using the BIC as compared with the AIC. Again, a lower value of λ_B indicates a more preferable model and nested models are not required.

1.6.6 False Discovery Rate

If a large number of simultaneous (independent) hypothesis tests are conducted without multiple comparison adjustment then we would expect $100\alpha\%$ of the tests to be significant even if H_0 is true at the $100\alpha\%$ significance level.

Benjamini and Hochberg (1995) suggested a method to reduce the number of incorrectly classified significant results. Their method selects significant results by considering their p-values. When testing m hypotheses there are four categories that the result can fall into as shown in table 1.4.

	declared not significant	declared significant	total
true null hypotheses	U	T	n
false null hypotheses	N	S	m-n
total	m-R	R	m

Table 1.4: The 4 ways of classifying observations

The proportion of errors committed by falsely rejecting null hypotheses is $Q = T/(T + S)$. The false discovery rate (FDR) is defined to be the expectation of Q. The Benjamini and Hochberg (1995) algorithm is described below:

- Consider m null hypotheses $H_1, ..., H_m$ and their respective p-values $p_1, ..., p_m$.

- Arrange the p-values in ascending order so that $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$.
- Choose a value q^* to be the FDR.
- Compare each $p_{(i)}$ to $\frac{iq^*}{m}$. Let k be the smallest i such that $p_{(i)} > \frac{iq^*}{m}$.
- Reject all the null hypotheses with p-values $p_{(1)}, ..., p_{(k-1)}$ and accept all the others.

Benjamini and Hochberg (2000) refined this algorithm. The new algorithm is similar to the previous one in that it uses the p-values to select significant results, however, it uses the rate of change in the p-values to decide where the cut-off point should be instead of just the actual p-values themselves. Using either of these two algorithms results in the threshold for significance being increased and thus fewer results are deemed significant.

In this thesis we use the original 1995 method to identify significance when testing multiple hypotheses in chapters 3 and 4.

Chapter 2

Classification of Proteomic Spectra using a Deterministic Peak Finding Algorithm

2.1 Introduction

In section 1.4 an overview of the two main data reduction methods and the different methods of classification were presented. This chapter will consider the use of these methods to classify new spectra once the peaks in the data have been obtained. After a short motivation, the algorithm for identifying the peaks is described in section 2.3 and the results obtained are described in sections 2.4 and 2.5.

2.2 Motivation

When searching for biomarkers for disease in proteomic data the aim is to find a small subset of the available m/z values which correctly classify disease state. It would also be beneficial if these biomarkers had some biological relevance.

Preliminary work carried out using principal components analysis (PCA) has enabled mostly correct classification of the spectra in both the breast cancer and melanoma datasets. However, although PCA is successful at reducing the dimension of the datasets, subsequent classification has been based on the PCs obtained. These PCs are constructed using all of the data and thus are not very biologically interpretable. Results obtained from this preliminary analysis are presented with the main results of this chapter in sections 2.4 and 2.5.

In figure 2.1 it can be seen that a mass spectrum consists of a series of distinct peaks. Each peak is centred around a particular *m/z* value and has a height which can differ across spectra. This height represents abundance of molecules - a larger peak implies a larger number of molecules with that *m/z* value were identified in the sample. To classify spectra into groups we can use this differing peak height information.

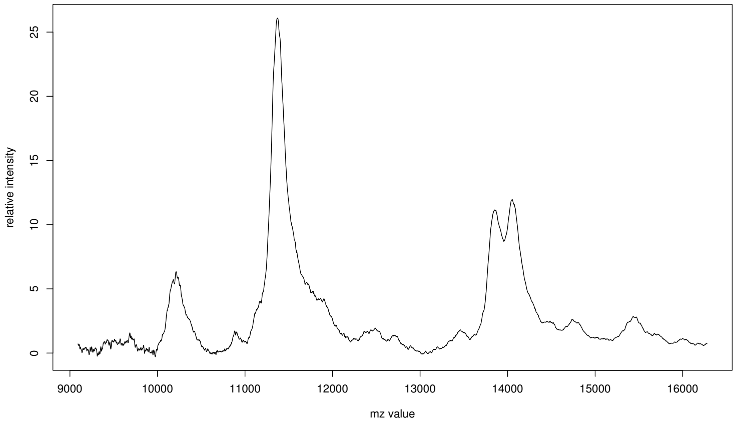

Figure 2.1: Example showing the presence of distinct peaks in a small section of a mass spectrum.

A method is now introduced by which spectra can be classified according to the peak locations and heights. This method provides much more interpretable reasons for classifications into particular groups.

2.3 The Peak Finding Method

Assume we have n spectra with k observations (m/z values) in each. Let Y_{ij} be the observed intensity at the i^{th} m/z value in the j^{th} spectrum. The aim is to model the peaks in the dataset ensuring common peak locations across all the spectra in order to aid the interpretability of the results.

We now outline a method for peak finding. To find the location of the first peak, find a location i_1 such that $i_1 = \text{argmax}_i \sum_{j=1}^{n} Y_{ij}$. This is equivalent to finding the largest peak in the mean spectrum. Place a Gaussian kernel of the form $c_{1j}N(\mu_1, \sigma_1^2) = c_{1j}f_1$ at this location. As suggested by the chemistry (e.g. Hortin, 2006), we model σ_i to be proportional to μ_i, the peak location, remembering that the constant of proportionality ξ still needs to be chosen. The scaling parameter c_{1j} then needs to be calculated for each spectrum j in order that the height of the fitted peak matches the height of the data at that point. In order to find the location of subsequent peaks, first form

$$X_{ij} = Y_{ij} - \sum_{p=1}^{P} c_{pj} f_p(i)$$

where P is the number of peaks already fitted. This effectively 'subtracts' the peaks already accounted for in the method. Then find an i_j such that $i_j = \text{argmax}_i \sum_{j=1}^{n} X_{ij}$, set $\mu_j = i_j$ and find the scaling parameters as before. Note that as we are only considering positive peak heights it is simpler to sum the X_{ij} instead of their squares. Repeat this algorithm until the desired number of peaks is fitted. We thus have a two-step procedure - we iteratively find the best position to fit a peak (conditional on the presence of existing peaks) and then calculate the required scaling parameters for the height of this peak in order that the fitted function matches the data exactly at the peak location.

We simplify matters in this model by assuming that ξ is a known constant, although a suitable value of ξ can be determined quite easily by trial and error. The algorithm takes of the order of seconds to find the peaks and so many values of ξ could be tested to find which appears to best match the data. However, in this thesis ξ is chosen by a least squares method minimising the residual error.

There is a potential problem with this method as it stands concerning peaks that overlap significantly. If all the peaks are far apart from each other then the small area in the tails of the distributions should not affect the heights of the other peaks significantly. However, if the peaks are close together then height contributions from the other peaks could affect how well the modelled peak heights actually fit the data. For an example when just two peaks are being fitted see figure 2.2.

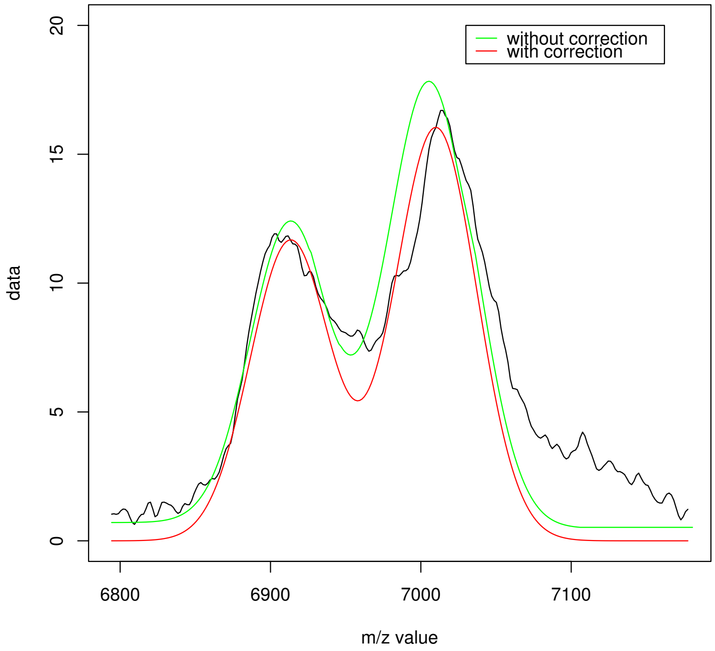

Figure 2.2: Example showing a peak arrangement with extra contribution to peak heights after a second peak has been fitted.

In figure 2.2, the peak around *m/z* value 7,000 is fitted according to the above algorithm. The green line shows the resulting fitted heights at each *m/z* value when the second peak is fitted. It is seen that there is an extra contribution to the peak height at, for example, point $\mu_1 = 7,000$. Also the previously correctly fitted peak at $m/z = 7,000$ contributes some extra height to peak 2 at, for example, point $\mu_2 = 6,900$. This occurs at all points where the distributions of the fitted peaks overlap and could result in substantially incorrect fitted heights. This phenomenon gets worse the larger the number of fitted peaks. In order to eliminate this problem we need to consider all peaks simultaneously, not separately as suggested above. In fitting the second peak we wish to match the heights of both peak 1 and peak 2. This can be achieved by solving the following two simultaneous equations in the two unknowns c_{1j} and c_{2j}:

$$\begin{aligned} c_{1j} f_1(i_1) + c_{2j} f_2(i_1) &= Y_{i_1,j} \\ c_{1j} f_1(i_2) + c_{2j} f_2(i_2) &= Y_{i_2,j}, \end{aligned} \tag{2.1}$$

where j is the spectrum number and i_p is the *m/z* value location of the p^{th} peak. This procedure is repeated every time a new peak is added, each time solving p equations in p unknowns for each spectrum. This system of equations can be solved in matrix notation by $C = F^{-1}Y$ where $C = [c_{1j} \;\ldots\; c_{pj}]$, $Y = [y_{i_1,j} \;\ldots\; y_{i_p,j}]$ and

$$F = \begin{bmatrix} f_1(i_1) & f_2(i_1) & \ldots & f_p(i_1) \\ f_1(i_2) & f_2(i_2) & \ldots & f_p(i_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(i_p) & f_2(i_p) & \ldots & f_p(i_p) \end{bmatrix}.$$

The red line in figure 2.2 shows the fitted heights at each *m/z* value when this approach has been used. The heights obtained are closer to the data than previously.

The algorithm can be adapted to either end after a specified number of peaks have been fitted or to continue until the X_{ij}'s are all less than a certain tolerance value.

In summary, the peak finding algorithm is as follows

1. For each m/z value, add up the intensities for each spectrum in the dataset at that m/z value to obtain a single overall spectrum.
2. Find the m/z value location of the largest peak in the overall spectrum.
3. For each spectrum, place a Gaussian kernel at this m/z value with standard deviation proportional to its mean.
4. (a) *If you are fitting the first peak:* For each spectrum, calculate the scaling parameter which will match the height of the Gaussian kernel to the data at that m/z value.

   (b) *If you are fitting subsequent peaks:* For each spectrum, calculate the scaling parameters for all peaks currently identified by the algorithm by solving simultaneous equations.
5. For each spectrum, subtract the modelled peak(s) from the original data.
6. As in step 1, sum the intensities in this subtracted dataset at each m/z value to obtain a new overall spectrum
7. Repeat steps 2 to 6 until a specified number of peaks have been fitted.

It should be noted that as we add the k^{th} peak the algorithm requires the inversion of n, $k \times k$ matrices. This results in fast calculations for small (< 50) numbers of peaks but results for a larger number of peaks are more computationally expensive. The algorithm is now applied to the breast cancer and melanoma datasets.

2.4 Applying the Algorithm to the Breast Cancer Dataset

2.4.1 Goodness of Fit

The algorithm was firstly applied to the whole dataset in order to assess if the fitted curves were a suitable approximation to the real data. Figure 2.3 shows the actual spectra and the modelled peaks based on the locations obtained using the peak finding algorithm for a small section of the data in the *adcon* group between 6,800 Da and 8,400 Da. One hundred and fifty peaks were used for the whole dataset and 8 of them were in this range. Peak locations were common across all spectra in the dataset. The black lines are the modelled peaks and the red lines are the cell line data which are reflected in the x axis.

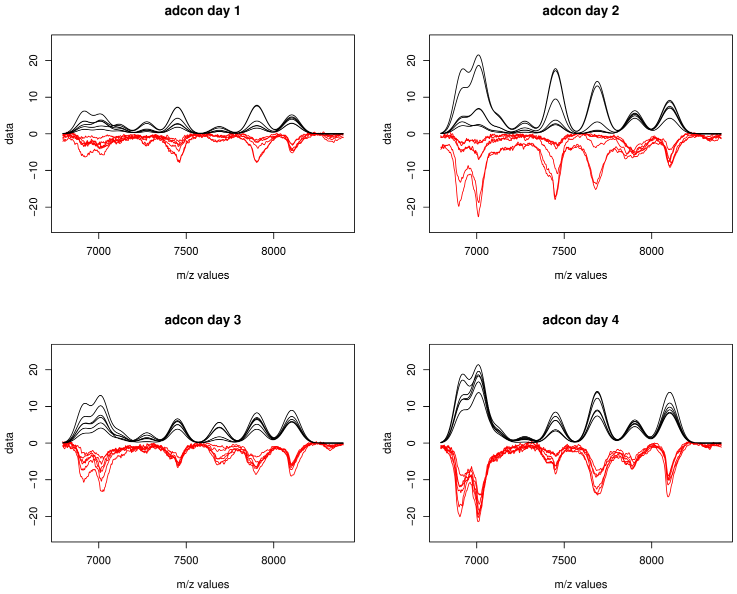

Figure 2.3: The peaks selected in the *adcon* group using the deterministic peak finding method and fitting 8 peaks. Red lines show the original data (reflected in the x-axis) and black lines show the fitted values using the peak finding algorithm.

It is clear that the algorithm is picking out the peak locations well. The differences in height for different spectra are apparent and are a good match to the original data. To analyse how well the model fits the data we now consider residual plots. These show the differences between the data and the fitted model and are presented in figure 2.4 for the *adcon* group over the same range of *m/z* values as previously.

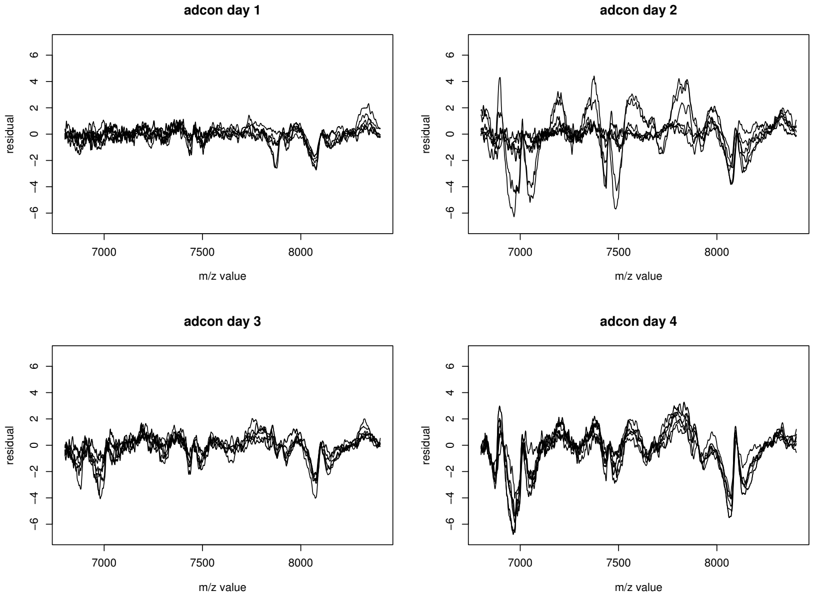

Figure 2.4: The residuals obtained for the *adcon* group by subtracting the model from the data.

Overall the residuals seem to be fairly small on day 1, however, there are some patterns evident in figure 2.4. There are two main reasons for these patterns. Firstly some of the peaks have too large a variance as can be seen at *m/z* values around 7,000 and 8,100 Da in figure 2.5. For the single peak at 8,100 Da the large variance results in a much wider peak than is actually present. Around 7,000 Da the presence of two close peaks each with too large a variance results in the trough between them being incorrectly modelled. Secondly, the prerequisite that the peaks have to have common locations across

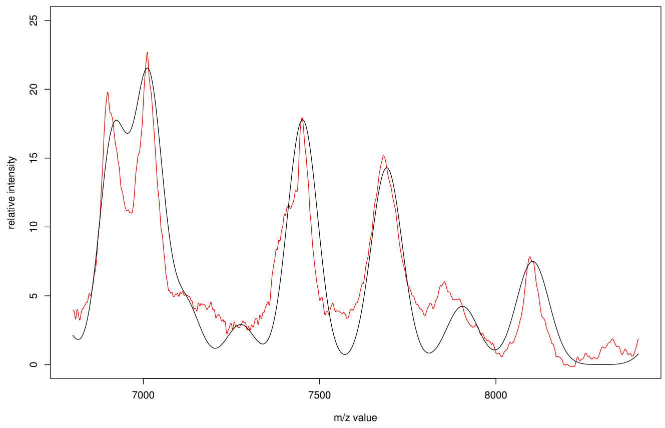

Figure 2.5: An example spectrum over the range 6,800-8,400 Da and the fitted model obtained from the peak finding algorithm. The red line shows one spectrum from the breast cancer dataset and the black line shows the fitted model.

all spectra means that, in some cases, the model is not matching to the correct peak height for that particular spectrum. For example, at 7,900 Da the model is matching to a datapoint on the slope of the peak.

Peak variances are linked to the m/z value at which they occur - the standard deviation is proportional to the location. So to ensure the peaks have smaller variances the constant of proportionality, ξ, can be made smaller. This, however, will have effects on all of the other peaks and may create less well fitting peaks somewhere else.

2.4.2 Classification

The complete dataset (144 spectra - 24 in each of 6 groups) was split into training and test sets with 16 spectra from each of the 6 groups comprising the training set - 96 spectra in total. The peak finding procedure detailed in the previous section was run on the training set using $\xi = 0.000039$ for the value of the proportionality constant. This value was obtained from the MCMC analysis to be presented in chapter 3. A selection of ξ values were tested around the value 0.000039 and the resulting peak locations did not change substantially. The peak locations obtained were used to fit the model to the test data and then various classification methods were employed to predict the group memberships of the test spectra. This algorithm was repeated 1,000 times.

Three different sets of methods for classifying the test data were used - support vector machines (SVM), discriminant analysis (linear and quadratic, LDA and QDA) and K-nearest-neighbour (KNN, $K = 1$, 5, 10 or 20) all of which are described in more detail in section 1.4.2. The peak finding algorithm fitted the peaks in order of size and, since it is not necessarily true that the larger peaks are the better classifiers, the peak locations obtained needed to be reordered. The most 'dataset independent' way of doing this was to order the peaks by best classification of the training data using the same method. For each method of classification the best 50 classifiers were used to classify the data instead of the full 150 as this greatly reduced the computation time. Six of the methods of classification thus required the peaks to be reordered. It is impossible to order the peaks by best classification of the training data when using 1-nearest-neighbour as the training set is always perfectly classified. The classification results are summarised in table 2.1 which shows the method used to classify, the maximum percentage of correct classifications with a 4 s.d. range and the number of peaks that had to be included to reach this maximum percentage. PCA and ICA were also carried out on the complete spectra before classification. As previously the PCs/ICs were ordered by best classification of the training data before their use to classify a test set. A summary of the results of this analysis is shown in table 2.2. The complete classification curves for all analyses are shown in figures 2.6 and 2.7.

method	correctly classified	± 2s.d.	no of peaks
LDA	84%	(78,90)	28
SVM	75%	(69,81)	50
KNN (K=5)	74%	(67,81)	22
KNN (K=10)	72%	(64,80)	19
KNN (K=1)	72%	(65,79)	44
QDA	70%	(62,78)	6
KNN (K=20)	70%	(62,78)	21

Table 2.1: Percentage of correct classifications obtained using the peak finding algorithm and 7 different classification methods.

method	correctly classified	± 2s.d.	no of pcs/ics
PCA+LDA	89%	(84,94)	50
PCA+SVM	83%	(78,88)	30
PCA+QDA	76%	(70,82)	9
PCA+KNN (K=1)	71%	(64,77)	50
PCA+KNN (K=5)	65%	(58,72)	50
PCA+KNN (K=10)	62%	(54,70)	50
PCA+KNN (K=20)	59%	(51,67)	50
ICA+LDA	90%	(85,95)	50
ICA+SVM	84%	(79,89)	30
ICA+KNN (K=5)	75%	(68,82)	25
ICA+KNN (K=10)	74%	(68,80)	19
ICA+KNN (K=1)	74%	(68,80)	50
ICA+KNN (K=20)	72%	(64,78)	20
ICA+QDA	41%	(32,50)	10

Table 2.2: Percentage of correct classifications obtained using PCA/ICA on the complete spectra and 7 different classification methods.

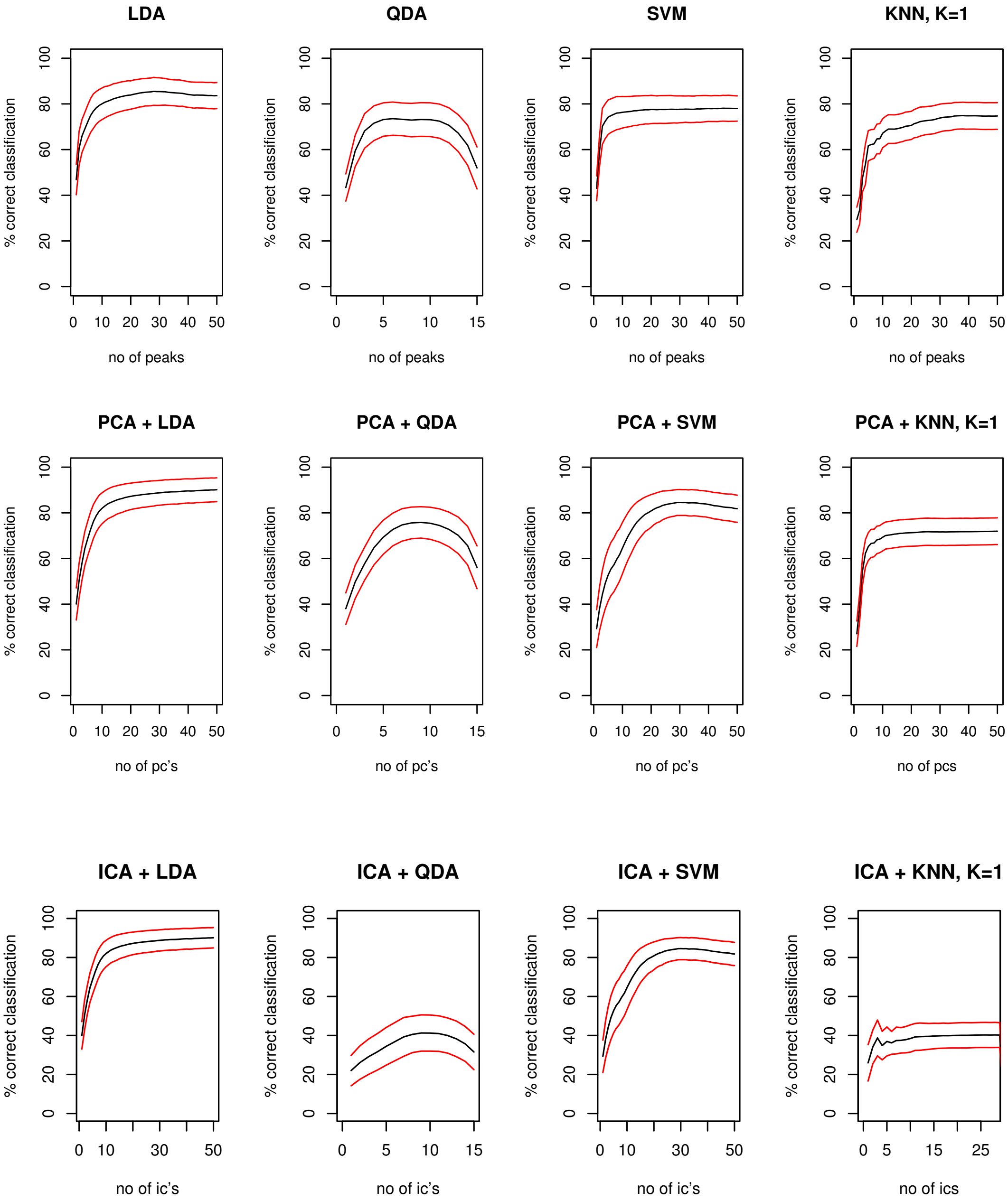

Figure 2.6: Classification curves using LDA, QDA, SVM and KNN ($K = 1$) combined with the fitted peak heights, and the PCs and ICs obtained from the entire dataset. Black lines show the mean and red lines ± 2 s.d.

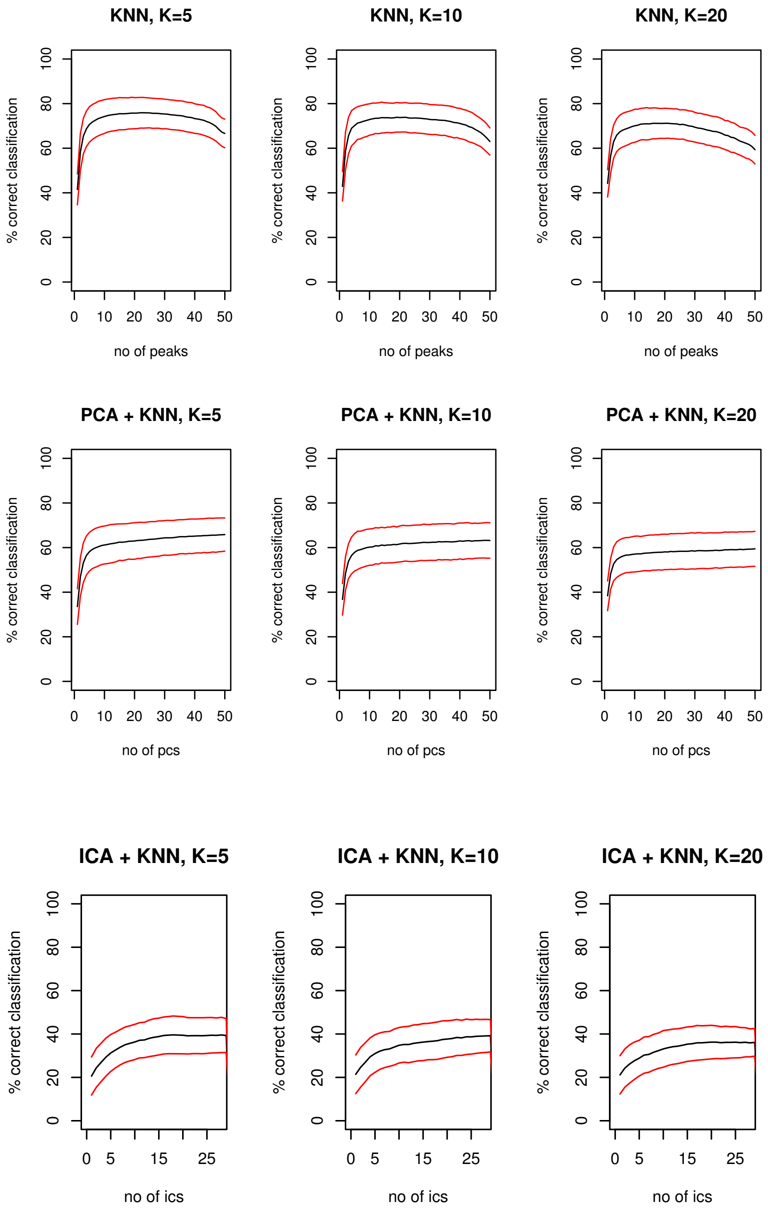

Figure 2.7: Classification curves using KNN ($K = 5, 10, 20$) combined with the fitted peak heights, and the PCs and ICs obtained from the entire dataset. Black lines show the mean and red lines ± 2 s.d.

From tables 2.1 and 2.2 it can be seen that LDA and SVM appear to be the better classifiers achieving correct classification rates of 84%-90% and 75%-84% respectively. Note that although the PCA/ICA classification rates were calculated using the complete spectra and are thus relatively high, the correct classification rates for the peak finding method are not much lower despite the reduction in the amount of data used. Also the PCA classifications require around 50 PCs to obtain their maximum correct classification rates, whilst the peak finding algorithm often requires fewer peaks to reach its maximum.

Using KNN on the PCs results in a lower number of correct classifications than using KNN on the peaks for all 4 values of K considered. Using KNN on the ICs results in a similar classification rate to the peak finding results. QDA does particularly badly when combined with ICA resulting in only 41% correct classifications.

Tables 2.3 and 2.4 show the peak locations most often identified as the best classifiers of the training data, when using LDA and SVM respectively, over all 1,000 iterations. The m/z values of 8,104 Da, 7,453 Da and 4,393 Da are the best first classifiers of the training data using both classification methods collectively being chosen first 96.6% and 90.1% of the time for LDA and SVM respectively. Note that only classifiers with totals above 150 are shown.

In figures 2.8 to 2.10 plots of the data are shown around the three m/z values in tables 2.3 and 2.4 which are selected most often as the first best classifier. It is clear why these locations were picked as the best classifiers of the data as the differences in height between the groups are so large. For the peak in figure 2.8 at an m/z value of 8,104 Da the distinction between classes is quite clear. The *adcon* and *adtax* groups have the highest peaks, the *tdcon* and *tdtax* groups have middle sized peaks and the *mccon* and *mctax* groups have flatter peaks.

	order of best classification of training data						
peak location	1st	2nd	3rd	4th	5th	6th	total
8,104	664	129	8	10	12	9	832
7,453	180	384	14	12	9	11	610
11,701	0	1	265	97	47	35	445
5,416	0	89	79	59	93	73	393
11,133	0	9	104	129	64	48	354
4,393	122	6	53	36	61	41	319
10,231	7	47	34	65	50	43	246
4,345	0	70	43	42	28	37	220
4,881	18	32	42	46	26	27	191
13,835	0	0	16	45	49	72	182
3,050	0	88	7	29	19	10	153
total	991	855	665	570	458	406	

Table 2.3: Peak locations which best classify the training data using LDA.

	order of best classification of training data						
peak location	1st	2nd	3rd	4th	5th	6th	total
4,393	243	83	209	66	47	38	686
8,104	390	112	71	27	24	17	641
4,345	9	267	143	87	62	38	606
5,416	0	102	203	59	44	21	429
4,881	59	150	52	25	44	52	382
7,453	268	34	14	16	19	13	364
2,180	0	0	8	67	95	66	236
10,231	5	62	22	27	35	28	179
5,707	0	27	27	38	45	38	175
5,366	0	25	27	60	23	21	156
total	974	862	776	472	438	331	

Table 2.4: Peak locations which best classify the training data using SVM.

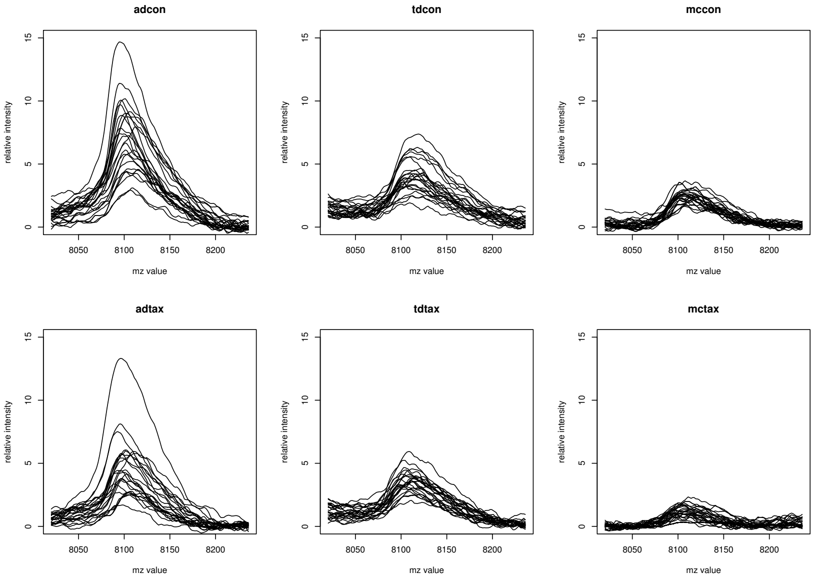

Figure 2.8: Plot of the data around m/z value 8,104 separated by group.

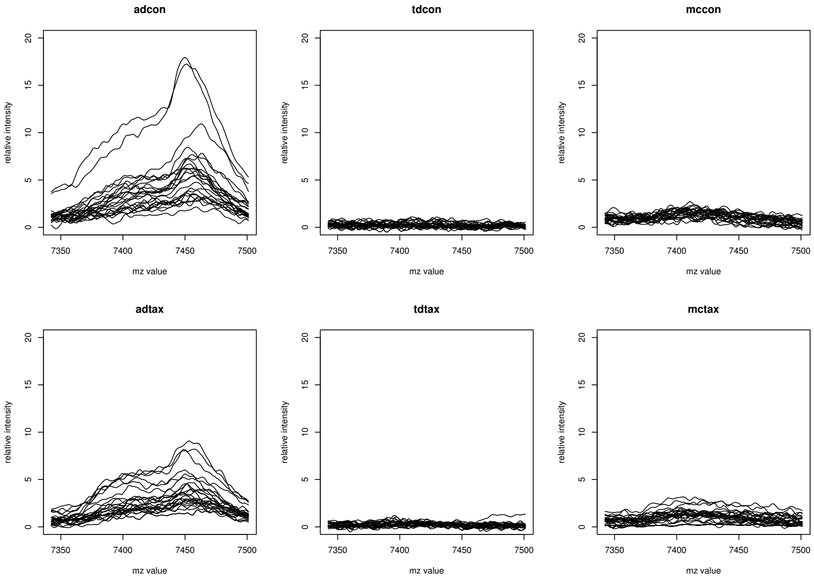

Figure 2.9: Plot of the data around m/z value 7,453 separated by group.

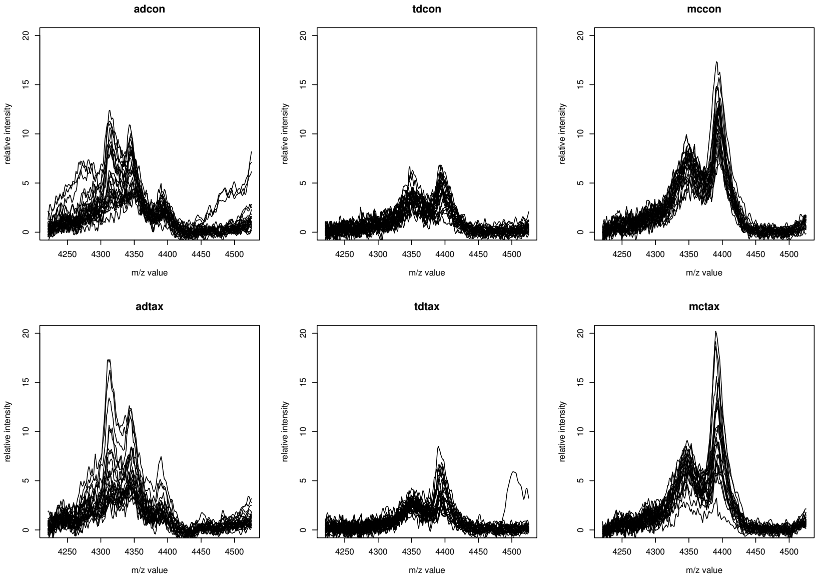

Figure 2.10: Plot of the data around *m/z* value 4,393 separated by group.

The peak in figure 2.9 at an *m/z* value of 7,453 Da is the next best classifier. For this peak we can see that the heights in the *adcon* and *adtax* groups are much higher than the other four groups. For both the peak at 8,104 Da and the peak at 7,453 Da the heights for the control spectra are larger than those for the respective Taxol treated spectra. This suggests that the Taxol treatment affects a cell by reducing the number of molecules at these *m/z* values. For the peak in figure 2.10 at an *m/z* value of 4,393 Da the opposite is true. The peak heights are higher in the treated spectra than in the controls. The largest peaks at this *m/z* value are to be found in the *mccon* and *mctax* spectra. The heights in the other four groups are comparable in size.

Tables 2.5 to 2.6 show misclassification tables based on 1,000 simulations for the LDA and SVM classification methods at the optimum number of peak locations (indicated in column 4 of table 2.1). It appears that it is relatively easy to discriminate between the three different types of cell-line but it is much harder to differentiate between Taxol-treated or non-Taxol-treated. The latter is the more important comparison for the physician.

		classified as					
		adcon	adtax	tdcon	tdtax	mccon	mctax
true group	adcon	79.2	15.3	0.1	0	5.2	0.3
	adtax	10.5	85.8	0	0	3.4	0.3
	tdcon	0	0.8	88.9	9.7	0.2	0.4
	tdtax	0	0.1	10.0	87.7	0.4	1.8
	mccon	0.2	0.1	0	0	88.9	10.8
	mctax	0.3	0.1	0	0	17.2	82.4

Table 2.5: Percentage of correct classifications using LDA on the fitted peak heights at 30 locations.

		classified as					
		adcon	adtax	tdcon	tdtax	mccon	mctax
true group	adcon	67.4	24.6	0	0	8.0	0
	adtax	19.6	75.7	0	0	4.8	0
	tdcon	0.1	0	94.3	2.9	2.6	0
	tdtax	4.4	0	12.0	80.7	2.8	0.2
	mccon	2.1	0.7	0	0	80.7	16.5
	mctax	2.0	2.2	0	0	26.2	69.6

Table 2.6: Percentage of correct classifications using SVM on the fitted peak heights at 50 locations.

There are an appreciable amount of *adcon* and *adtax* spectra misclassified into the *mccon* group. This is not wholly unexpected as ADR/MCF-7 cell-lines are treated versions of MCF-7 cell-lines and thus some similarities should be expected.

2.5 Applying the Algorithm to the Melanoma Dataset

2.5.1 Goodness of Fit

To check if the algorithm was giving a suitable approximation to the data the whole dataset was modelled and a fitted curve for each spectrum obtained using the deterministic peak finding algorithm. In figure 2.11 we see a small subset of the data, between 6,800 Da and 8,400 Da which shows sections from six actual spectra and the modelled peaks based on the locations obtained using the peak finding algorithm. One hundred and fifty peaks were used for the whole dataset and 10 of them were in this range. The black lines are the modelled peaks and the red lines are the cell line data which are reflected in the x axis.

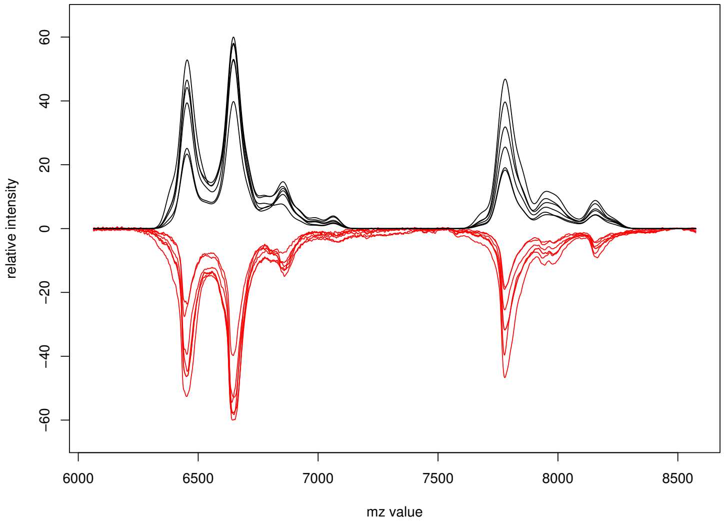

Figure 2.11: The peaks selected in 6 spectra from the melanoma dataset using the deterministic peak finding method and fitting 10 peaks.

The peaks in the data are identified well by the algorithm and height differences between different spectra are apparent which results in a good match to the original data. We now consider residual plots to assess how well the model fits the data. Residual plots for the same six spectra shown in figure 2.11 are presented in figure 2.12.

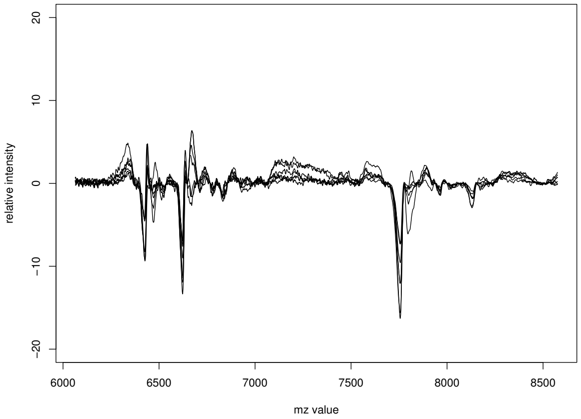

Figure 2.12: The residuals obtained for the 6 spectra in figure 2.11 by subtracting the model from the data.

In general the residuals are fairly small, however, there are three areas in figure 2.12 where the residuals show patterns. As discussed in section 2.4.1 for the breast cancer dataset two possible reasons are peaks having too large a variance or modelled peaks not having exactly the same location as the data due to the common location restriction. These reasons explain the residual patterns around the *m/z* value 6,500 Da. In addition we note that for the melanoma dataset many of the peaks appear to have a non-Gaussian shape. This results in modelled peaks with broader slopes than the data on one side of the peak and thus the residuals show this pattern. From figure 2.13 we can see that the modelled peak at an *m/z* value of 7,800 Da gives rise to this type of residual pattern. Some of these problems will be addressed in chapter 3 by using multiple peaks.

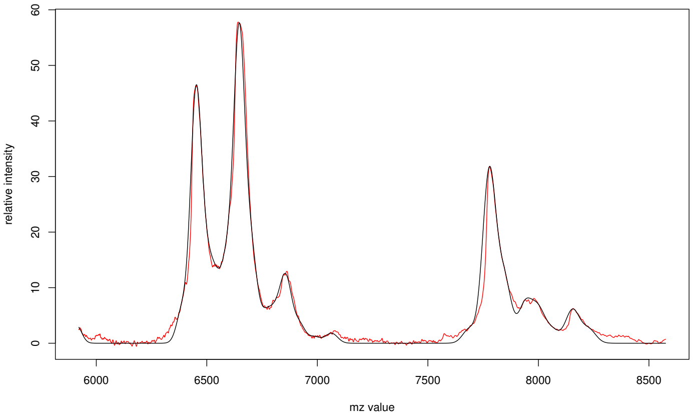

Figure 2.13: An example spectrum over the range 6,000-8,500 Da and the fitted model from the peak finding algorithm. The red line shows one spectrum from the melanoma dataset and the black line shows the fitted model.

2.5.2 Classification

All of the spectra in the dataset (205 spectra - 101 in stage I and 104 in stage IV) were designated 'training' or 'test' with 68 spectra from each of the stages comprising the training set. Using the peak finding algorithm with $\xi = 0.000020$, obtained from the MCMC analysis of this dataset to be presented in chapter 3, the training spectra were modelled. The peak locations obtained were used to fit the model to the test data and then the various classification methods used previously were used to predict the stages of the test spectra. The peak locations were reordered according to best classification of the training data and, for LDA/QDA, by the absolute t-statistic at each peak. The classification results from 1,000 repetitions of this algorithm are summarised in table 2.7. PCA and ICA were also carried out on the complete spectra before classification. As previously the PCs/ICs were ordered by best classification of the training data before using the best 50 to classify the test data. A summary of the results of the PCA/ICA analyses are shown in table 2.8. The complete classification curves for all analyses are shown in figures 2.14 and 2.15.

method	correctly classified	± 2s.d.	no of peaks
t-test & LDA	85%	(80,90)	12
SVM	83%	(78,88)	48
bestclass & LDA	81%	(76,86)	26
KNN (K=20)	81%	(76,86)	28
t-test & QDA	80%	(75,85)	8
KNN (K=10)	80%	(75,85)	23
KNN (K=5)	80%	(75,85)	31
bestclass & QDA	78%	(74,82)	3
KNN (K=1)	75%	(70,80)	50

Table 2.7: Percentage of correct classifications obtained using the peak finding algorithm and 7 different classification methods.

method	correctly classified	± 2s.d.	no of pcs/ics
PCA+LDA	86%	(82,90)	48
PCA+SVM	85%	(80,90)	50
PCA+QDA	81%	(75,87)	28
PCA+KNN (K=1)	76%	(70,82)	50
PCA+KNN (K=20)	76%	(70,82)	50
PCA+KNN (K=10)	74%	(68,80)	50
PCA+KNN (K=5)	65%	(55,75)	50
ICA+QDA	89%	(80,98)	28
ICA+LDA	85%	(80,90)	30
ICA+SVM	82%	(76,88)	30
ICA+KNN (K=1)	60%	(38,82)	1
ICA+KNN (K=5)	60%	(38,82)	5
ICA+KNN (K=10)	60%	(41,79)	2
ICA+KNN (K=20)	60%	(41,79)	2

Table 2.8: Percentage of correct classifications obtained using PCA/ICA on the complete spectra and 7 different classification methods.

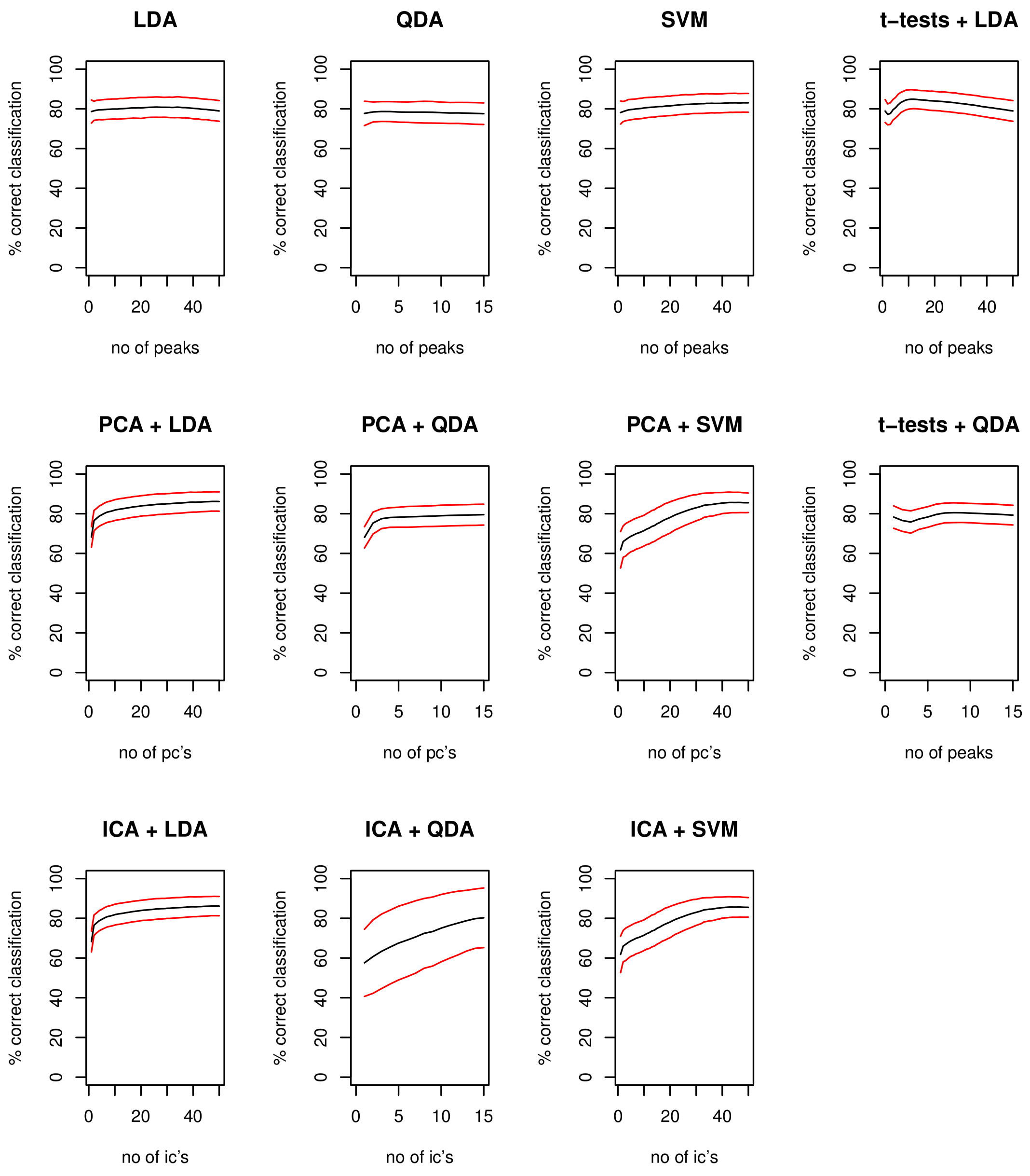

Figure 2.14: Classification curves using LDA, QDA and SVM combined with the fitted peak heights, and the PCs and ICs obtained from the entire dataset. Black lines show the mean and red lines ± 2 s.d.

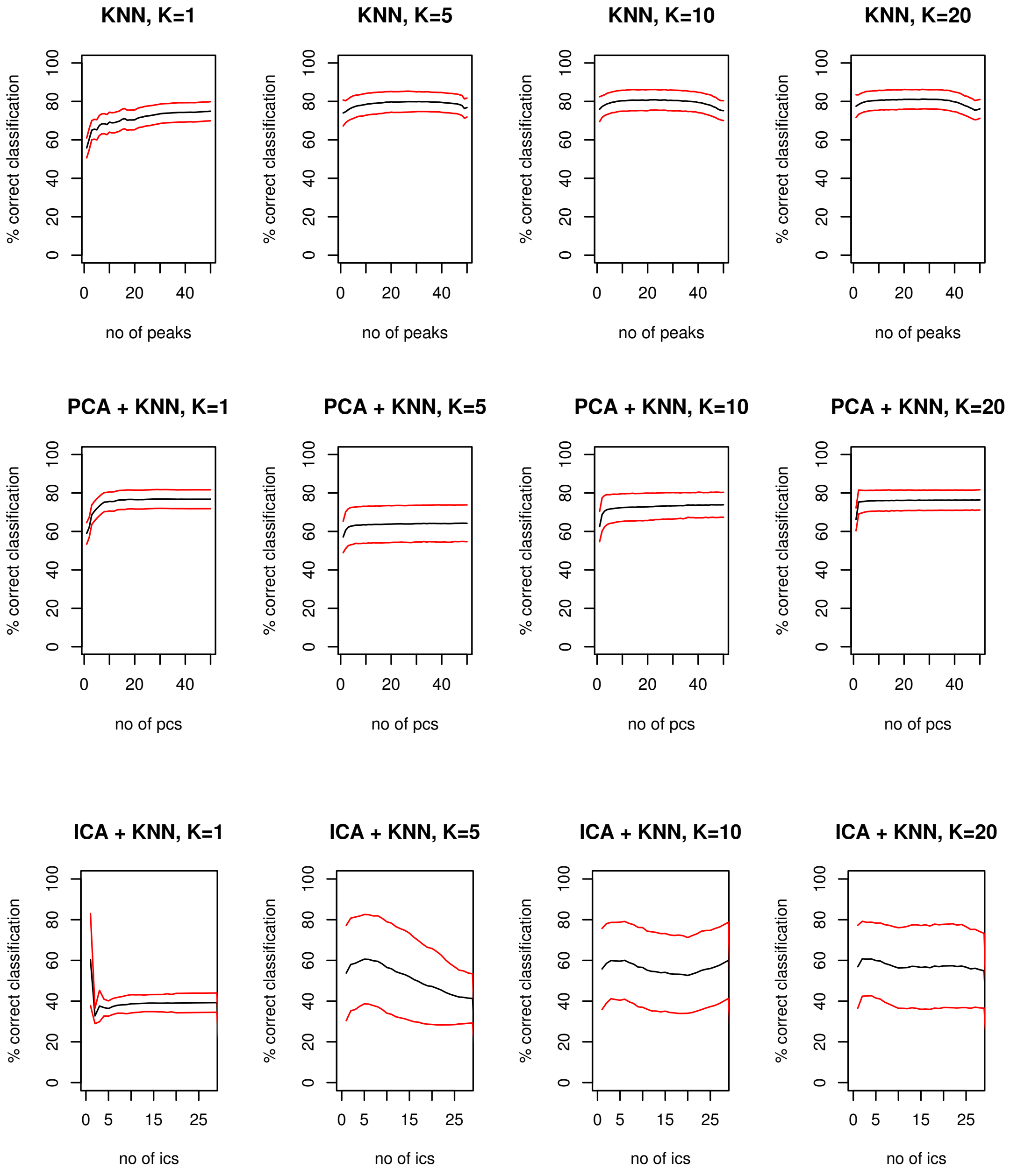

Figure 2.15: Classification curves using KNN ($K = 1, 5, 10, 20$) combined with the fitted peak heights, and the PCs and ICs obtained from the entire dataset. Black lines show the mean and red lines ± 2 s.d.

LDA and SVM are shown to be good classifiers in tables 2.7 and 2.8 with correct classification rates of 81%-86% and 82%-85% respectively. Note that, similarly to the breast cancer dataset, the reduction in data used for the peak finding method as compared with the complete spectra does not result in much lower correct classification rates. Again the PCA classifications require around 50 PCs to obtain their maximum correct classification rates whereas the peak finding algorithm often requires fewer peaks to reach its maximum. When the same data were analysed by Mian et al. (2005) the data were split into training, test and blind sets and classification was carried out using artificial neural networks. The correct classification rate using these methods was 88%.

Using KNN on the ICs results in a much lower number of correct classifications than any of the other methods with the average maximum correct classification being 60%. However, the standard deviations in each of the IC + KNN cases are large. Not only does this give unreliable results but also puts the lower confidence limit below 50% - suggesting results worse than the value expected by chance. For the case $K = 1$ the entire classification curve except for the first point is below the 50% level. ICA combined with QDA also results in a large standard deviation. Neither of these methods should be used to obtain reliable information about classifications.

Considering the results over all 1,000 iterations, the peak locations identified most often as the best classifiers of the training data are shown in tables 2.9 and 2.10 for the LDA and SVM classification methods. The best first classifiers of the data occur at m/z values of 3,885 Da, 28,160 Da, 3,316 Da and 2,227 Da using both LDA and SVM. They are collectively chosen first 99.5% and 99.2% of the time respectively. The three m/z values which are selected most often as the first best classifier in tables 2.9 and 2.10 are plotted in figures 2.16 to 2.18. In all three cases the average heights of the peaks in stage I appears to be higher than those in stage IV.

	order of best classification of training data						
peak location	1st	2nd	3rd	4th	5th	6th	total
3,885	917	51	1	0	1	1	971
28,160	60	153	43	24	13	17	310
8,154	0	2	9	51	69	62	193
2,040	0	108	22	21	15	23	189
7,978	0	4	23	41	62	46	176
3,316	12	22	82	24	17	16	173
8,949	0	3	12	48	52	50	165
8,820	0	108	25	15	7	8	163
2,771	0	69	54	17	16	5	161
2,227	6	67	37	19	17	15	161
7,777	0	2	31	28	49	50	160
2,495	2	92	30	11	7	9	151
total	997	681	369	299	325	302	

Table 2.9: Peak locations which best classify the training data using LDA.

	order of best classification of training data						
peak location	1st	2nd	3rd	4th	5th	6th	total
3,885	916	25	5	2	3	0	951
2,040	0	104	59	68	54	44	329
28,160	29	85	58	41	30	34	277
2,771	0	175	43	19	10	19	266
2,359	0	17	46	59	47	48	217
3,316	14	45	53	35	30	20	197
8,949	0	1	37	48	63	40	189
2,227	33	57	43	16	15	15	179
2,269	0	61	42	35	17	22	177
2,495	0	86	40	18	14	7	165
2,540	0	24	20	33	29	23	129
9,460	0	17	36	29	13	17	112
total	992	697	482	403	325	289	

Table 2.10: Peak locations which best classify the training data using SVM.

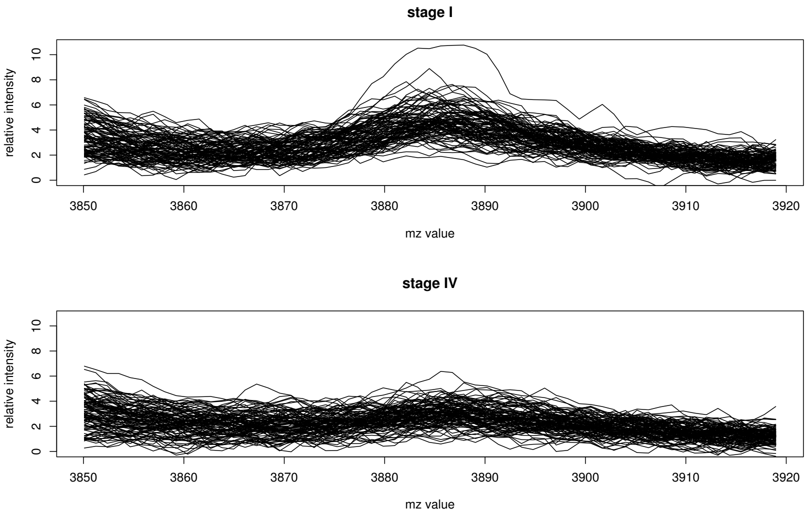

Figure 2.16: Plot of the data around *m/z* value 3,885 separated by stage.

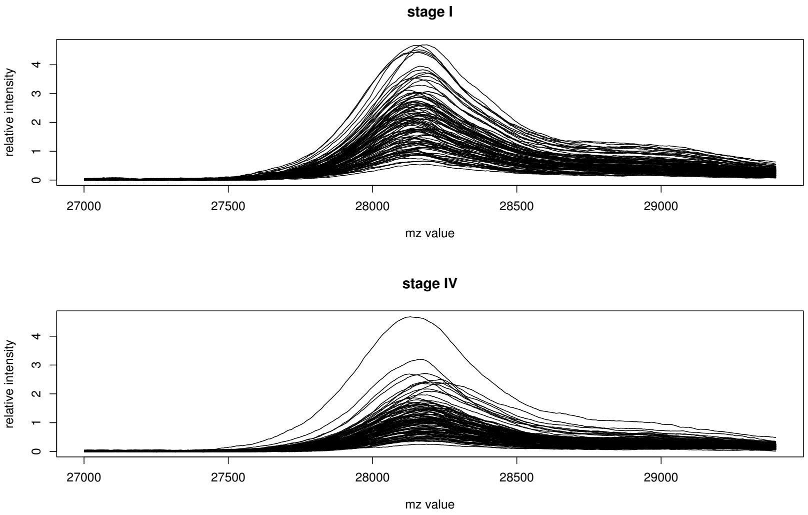

Figure 2.17: Plot of the data around *m/z* value 28,160 separated by stage.

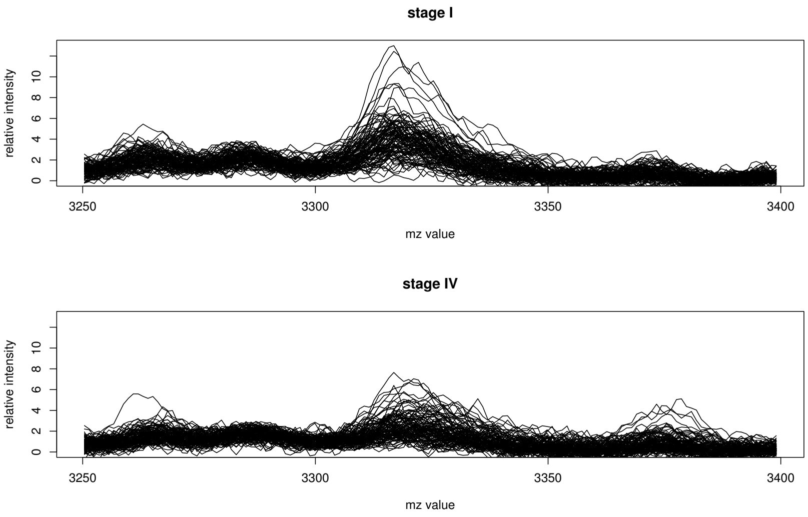

Figure 2.18: Plot of the data around m/z value 3,316 separated by stage.

A little further analysis shows why these three peaks are good classifiers. If we consider the relative intensities for all spectra at one peak location we can identify a dividing point at a particular relative intensity which results in the majority of stage I spectra having lower (higher) intensities and the majority of stage IV spectra having higher (lower) intensities.

For the peak in figure 2.16 at an m/z value of 3,885 Da, a dividing point is located at a relative intensity of 4. At this point 75% of the stage I spectra have higher intensities and 88% of the stage IV spectra have lower intensities. For the peak in figure 2.17 at an m/z value of 28,160 Da, the dividing point is at a relative intensity of 1.4. Then 73% of the stage I spectra have higher intensities and 68% of the stage IV spectra have lower intensities. Lastly for the peak in figure 2.18 at an m/z value of 3,316 Da, a dividing point is located at a relative intensity of 3.1. At this point 70% of the stage I spectra have higher intensities and 80% of the stage IV spectra have lower intensities.

Tables 2.11 to 2.12 show misclassification tables for the LDA and SVM classification methods at the optimum number of peak locations (indicated in column 4 of table 2.7). The results for the two methods are similar although the percentage of correct stage I classifications is slightly higher when using SVM. To obtain this 4.2% increase, however, requires the use of nearly twice as many peak locations.

		classified as	
		stage 1	stage 4
true	stage 1	76.1	23.9
	stage 4	14.6	85.4

Table 2.11: Percentage of correct classifications using LDA on the fitted peak heights at 26 locations.

		classified as	
		stage 1	stage 4
true	stage 1	80.3	19.7
	stage 4	14.3	85.7

Table 2.12: Percentage of correct classifications using SVM on the fitted peak heights at 48 locations.

2.6 Summary

In this chapter we have developed a deterministic peak finding algorithm and have shown how it can be used to successfully model mass spectrometry data. Also it has been shown how we can use the models obtained to classify new spectra. The results were obtained using C++ and the *R* programming language.

The algorithm is good at identifying the peaks in the data and it provides differing heights for peaks which match the data well. Some differences in the peak heights between the data and the model are apparent but these can mostly be attributed to the restriction that was in place of common peak locations across spectra.

When considering classifying new spectra it was found that the method of classification could drastically change the percentage of correct classifications obtained. LDA and SVM provided consistently good results over both datasets. The SVM method provided better results than LDA for the two group classification in the melanoma dataset. This could be due to the simplicity of the problem when compared with the 6 group case in the breast cancer dataset or alternatively it could be due to the greater amount of training data available for this dataset. For the breast cancer dataset there were only 4 spectra from each day of each group in the training set which may not be a representative sample of the population of spectra. It is also possible that the SVM method performed better than the LDA method because it is more robust against distant observations on the wrong side of the boundary. The possibility of some outliers being present is taken into account when constructing an SVM and the classification boundary should not alter much, if at all, from the boundary if those outliers were excluded. However, in LDA the presence of an outlier would skew the discrimination boundary.

When using LDA and SVM to classify, the results obtained using the algorithm were comparable with using the information contained in the compete spectra despite the reduction in the number of available datapoints. This shows promise for the use of data reduction methods to analyse high-dimensional proteomic data.

Whichever method was used to classify new spectra in the breast cancer dataset it was found that it was relatively easy to distinguish between the three different types of spectrum (MCF-7/ADR, T47D or MCF-7). It was harder to separate out control and treated specimens of the same type of cell line. Some MCF-7/ADR spectra were misclassified into the MCF-7 group which should not be surprising as they are both derived from the same original cell-lines.

In Mian et al. (2003) the breast cancer dataset was studied using *artificial neural networks* (ANNs). This research highlighted the *m/z* values 10,518 Da, 11,100 Da, 11,687 Da and 13,239 Da as showing good classification ability between control and treated cell-lines. Only *m/z* values between 10 kDa and 15 kDa were considered in that analysis. The methods described in this chapter provide similar results to two of these values at 11,133 Da and 11,701 Da. These values are shown in tables 2.3 and 2.4 as some of the best classifiers of the training data.

In Mian et al. (2005) the melanoma dataset was also studied using ANNs. The research concluded that the best predictive capability came from the region between 2,000 Da and 5,000 Da and that very little predictive value was obtained from the range between 10,000 Da and 15,000 Da. These observations are replicated in the work presented in this chapter. If we consider tables 2.9 and 2.10 we see that the majority of locations which best classify the data are in the 2-5 kDa range and that no good classifiers exist between 10-15 kDa. Indeed the only classifier with an *m/z* value greater than 10kDa is observed at 28,160 Da. This location is one of the top three best classifiers identified in this chapter.

There are three main advantages to analysing spectra using the methods considered in this chapter. Firstly, the initial peak finding method relies on simple calculations which can be carried out quickly. The datasets considered in this chapter were of length $\approx$ 14,000. Fitting 150 peaks to one of the datasets whilst also correcting the heights for peaks close to one another took around 90 seconds. This is a much quicker method of identifying peaks than the MCMC analysis that will be carried out in chapter 3. Secondly, in the classification step the new spectra are only being classified using the peak heights at each location. This drastically reduces the dimension of the classification problem from 144 × 14,000 to 144 × 150 for the breast cancer dataset and from 205 × 14,000 to 205 × 150 for the melanoma. Lastly when we compare with traditional methods for data reduction, for example PCA, the peak finding algorithm provides us with much more interpretable reasons for classification into particular groups.

Chapter 3

Modelling Mass Spectrometry Data Using Markov Chain Monte Carlo Simulation

3.1 Introduction

In section 1.5 an overview of some basic Bayesian theory and a short introduction to the algorithms used in Markov Chain Monte Carlo (MCMC) was presented. This chapter will consider the use of these methods to model the available datasets. The initial model used in the MCMC simulations will be introduced in section 3.2 and the results obtained from this are described in section 3.3. More complex models are discussed in sections 3.4 and 3.5 and the results are compared with those already obtained.

3.2 Modelling the Data

3.2.1 The Model

The aim is to use mass spectra to firstly differentiate between drug-treated breast cancer cell-lines and non-treated controls as in Dryden et al. (2005), and Mian et al. (2003) who use neural networks; and secondly to differentiate between stage I and stage IV melanoma as in Mian et al. (2005).

It can be noted from figures 3.1 and 3.2 that the spectra consist of a sequence of peaks of varying heights. Figure 3.1 shows sections of 6 spectra in the breast cancer dataset, one from each of the groups, and figure 3.2 shows sections of two spectra from the melanoma dataset, one from each of the two stages. A possible modelling approach is therefore to fit a series of Gaussian peaks to the data with locations, heights and variances to be estimated. This approach can be implemented using the MCMC methods described in section 1.5 to construct samples from the joint posterior distribution of the unknown parameters, namely the locations, heights and variances of the peaks.

The model used for each datapoint y_{is} is distributed as $y_{is} \sim N(\theta_{is}, \tau^{-1})$ where the y_{is} are independent of each other and where θ_{is} is the sum of scaled Gaussian distributions:

$$\theta_{is} = \sum_{j=1}^{k} h_{js}(\xi\mu_j^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\xi\mu_j^2)^{-1}(x_i - \mu_j)^2\right) \tag{3.1}$$

where the index $i = 1, \ldots, p$ represents the position on the spectrum, x_i is the i^{th} m/z value, $s = 1, \ldots, n$ is the spectrum number and $j = 1, \ldots, k$ is the peak number. (The constant $1/(2\pi)$ has been subsumed in the h_{js} parameters.) The parameters μ_j and h_{js} represent respectively the location and the scaling that adjusts the height of the peaks in the model, and ξ is a constant of proportionality that models the fact that the standard deviation of the peaks increases linearly with the mean.

The means of the y_{is}'s are similar due to the model formulation, however, they are independent. The independence assumption on the y_{is}'s holds because the random errors are independent. There is no constraint that the integral of the whole spectrum remains fixed.

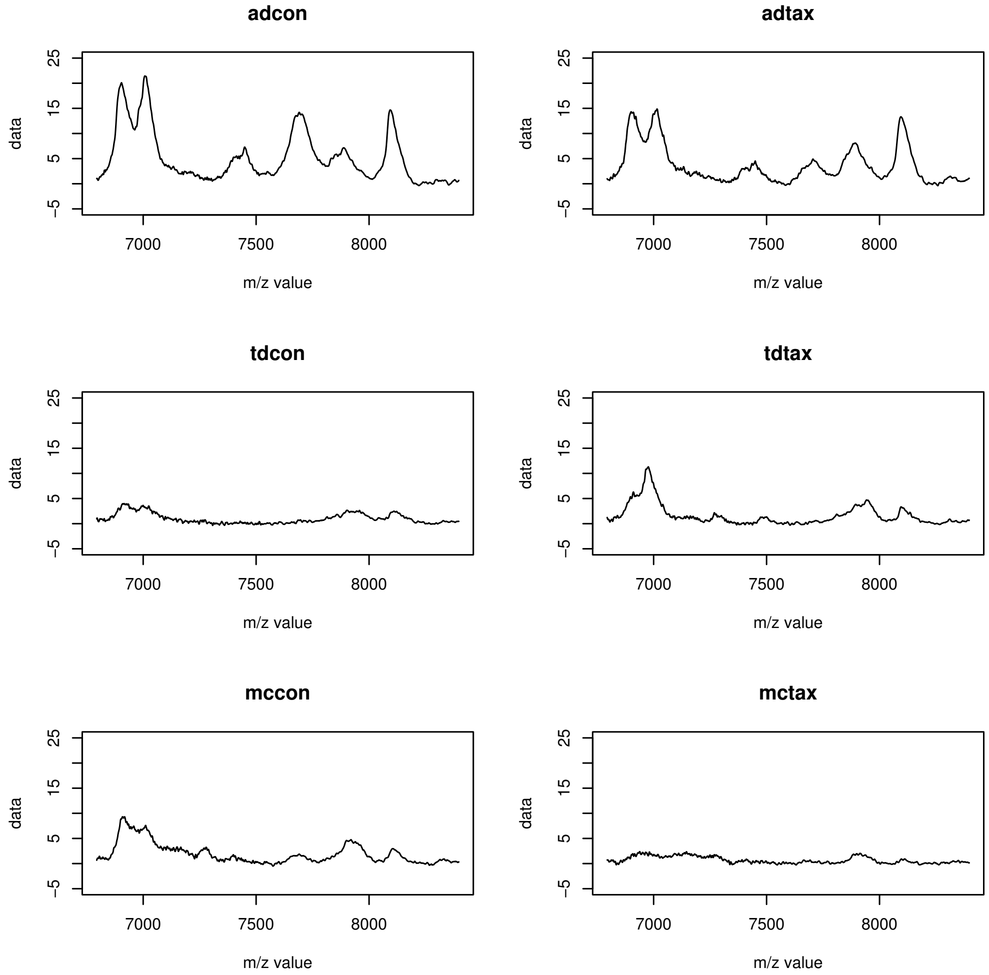

Figure 3.1: BREAST CANCER: Plots of the section of data between m/z values 6,800 and 8,400 for replicate 1 in each of the six groups on day 4.

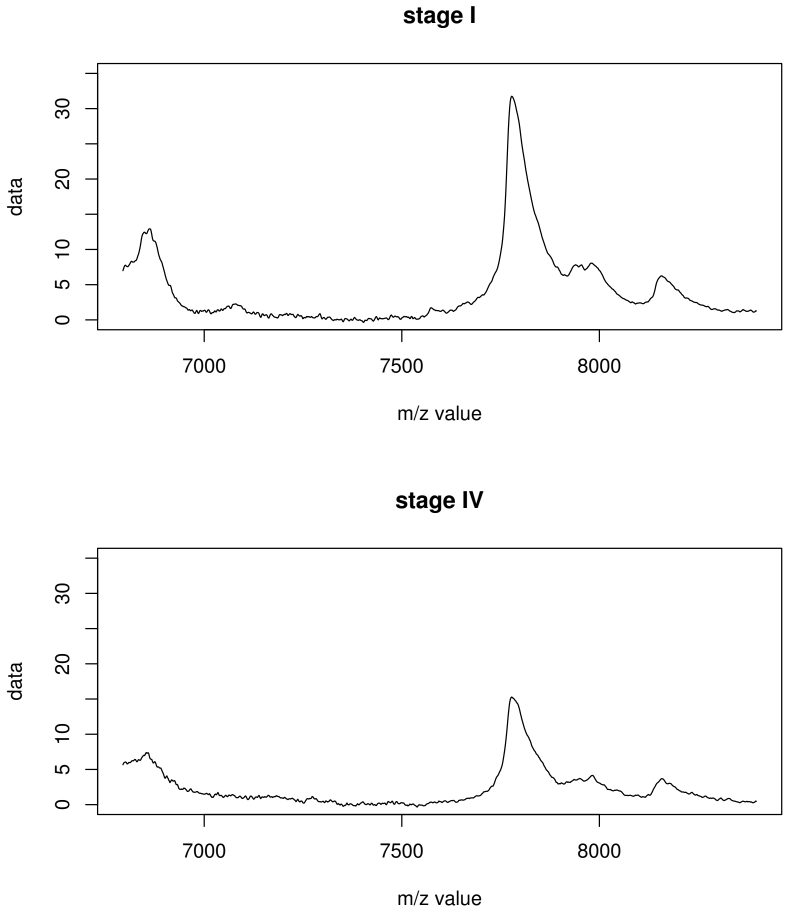

Figure 3.2: MELANOMA: Plots of the section of data between m/z values 6,800 and 8,400 for replicate 1 in each of the two stages of melanoma.

The model we wish to fit has common locations μ_j ($j = 1, \ldots, k$) and a common constant of proportionality ξ across the spectra. The heights of each peak are allowed to differ across spectra and the values of the h_{js}'s will indicate the presence or absence of a peak at a particular m/z value. Using common locations will enable us to determine the height difference between spectra at any particular location.

3.2.2 Priors for the Parameters

Variance parameters are restricted to be strictly positive values and so prior distributions such as the normal should not be used. The conjugate prior for a variance parameter with normal data is the inverse gamma distribution. This is equivalent to a gamma prior on the precision parameter $\tau = 1/\sigma^2$. We are considering vague priors for all the model parameters to try to ensure any inferences come from the actual data and not because of strong prior information. The vague gamma prior chosen for τ is $Gamma(\epsilon, \epsilon)$ with $\epsilon = 0.001$ which has mean $\frac{\epsilon}{\epsilon} = 1$ and large variance $\frac{\epsilon}{\epsilon^2} = \frac{1}{0.001} = 1000$.

The parameter ξ describes the proportionality of peak standard deviation to peak location. Since the peak standard deviation is constrained to be positive, the value of ξ must also be positive. A $Uniform(0, 0.01)$ distribution was used for the prior distribution in this case. The maximum possible peak location is around 30 kDa which makes $\xi \times \mu^2 = 0.01 \times 900 = 9$ and hence a maximum peak standard deviation of 3 kDa.

The height scaling parameters h are similarly constrained to be positive. However, plots of the data suggest that the heights are free to lie anywhere within a restricted range of values. A suitably noninformative distribution of this type is the uniform distribution and for this reason a $Uniform(0, 100)$ prior was chosen for these parameters.

When running the MCMC algorithms it is possible for the estimate of one location parameter to become very close to another. If this continues it is possible to have many parameters referring to the same peak and, in the extreme case, all location parameters could become equal which would not give biologically interpretable results. To ensure location parameters do not become too close without good reason a Strauss prior is used (Kelly and Ripley, 1976). A Strauss prior has two parameters: an intensity β and a tolerance R

and the joint distribution is:

$$\pi(\mu_1, \ldots, \mu_k) \propto e^{-\beta(\text{number of pairs of } \mu_j' s \text{ that are } R\text{-close})}. \tag{3.2}$$

Here *R-close* means within tolerance R. The tolerance R is how close peaks are allowed to be without penalty. The intensity β represents how strongly one wishes to penalise a 'close' proposal - the higher the value of β the higher the penalty. The use of this prior will penalise proposals where locations are too similar such that if a proposed peak is too close to another, it will only be accepted if it results in a higher posterior density. We use $\beta = 4,000$ and $R = 100$. This effectively ensures peaks cannot be within 100Da of each other unless the posterior density is greatly increased by close peaks.

These priors and the likelihood given by the data model in equation (3.1) are combined using Bayes' Theorem to give the following posterior density function

$$\begin{aligned} \textit{posterior} &\propto \pi(y|\xi, \mu, h, \tau)\pi(\xi)\pi(\tau)\pi(\mu)\pi(h) \\ &\propto \prod_{i=1}^{p}\prod_{s=1}^{n} \tau^{\frac{1}{2}} \exp\left(-\frac{\tau}{2}\left[y_{is} - \sum_{j=1}^{k} h_{js}(\xi\mu_j^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\xi\mu_j^2)^{-1}(x_i - \mu_j)^2\right)\right]^2\right) \\ &\times \quad \frac{\epsilon^{\epsilon}\tau^{\epsilon-1}e^{-\epsilon\tau}}{\Gamma(\epsilon)} \\ &\times \quad e^{-\beta(\text{ number of pairs of } \mu_j' s \text{ that are } R\text{-close})} \end{aligned} \tag{3.3}$$

where we assume that ξ, τ, μ and h are a priori independent. This is the distribution from which we wish to sample parameter values.

3.2.3 Updating the Parameters

Throughout this section $\lambda^{(t)}$ will be used to denote the set of all the current model parameters, and λ^* will be used to denote the proposed set of parameters. Since zero mean normal proposal distributions are used for all

Metropolis-Hastings updates their ratio is equal to 1 and does not appear in the acceptance probability calculations.

Updating τ

The complicated likelihood expression shown in equation (3.3) means the full conditional distributions of most of the parameters cannot be written down easily. However, the precision parameter τ is not involved in calculating θ_{is} and thus its full conditional distribution can be easily determined. Hence, the precision parameter τ is the only parameter in this model that can be updated using conjugate Gibbs sampling. The full conditional distribution of τ is

$$\begin{aligned}
&P(\tau|\mu_1, \ldots, \mu_k, h_{1,1}, \ldots, h_{ks}, \xi, y) \\
&\qquad\qquad \propto \quad \tau^{\frac{pn}{2}+\epsilon-1} \exp\left(-\tau\left[\frac{1}{2}\sum_{i=1}^{p}\sum_{s=1}^{n}(y_{is}-\theta_{is})^2+\epsilon\right]\right) \\
&\textit{i.e.} \qquad \tau \quad \sim \quad \text{Gamma}\left(\frac{pn}{2}+\epsilon \quad , \quad \frac{1}{2}\sum_{i=1}^{p}\sum_{s=1}^{n}(y_{is}-\theta_{is})^2+\epsilon\right). \qquad (3.4)
\end{aligned}$$

So at each iteration of the MCMC algorithms a new value for τ will be drawn from this distribution.

Updating ξ

When updating ξ the only part of the posterior that changes is the contribution from the likelihood. The prior contribution remains the same since the prior on ξ is uniform and thus is independent of the value of ξ. Assume that the current value of ξ is ξ_t and the proposed value is $\xi_* \sim N(\xi_t, \sigma_\xi^2)$. Then the equation for the acceptance probability α is:

$$\alpha(\lambda^{(t)}, \lambda^*) \quad = \quad \min\left\{1, \frac{\pi(y|\lambda^*)\pi(\lambda^*)}{\pi(y|\lambda^{(t)})\pi(\lambda^{(t)})}\right\}$$

$$= \quad \min\left\{1, \frac{\prod_{i=1}^{p}\prod_{s=1}^{n}\exp\left(-\frac{\tau}{2}\left[y_{is}-\sum_{j=1}^{k}\frac{h_{js}}{(\xi_*\mu_j^2)^{\frac{1}{2}}}\exp\left(-\frac{(x_i-\mu_j)^2}{2\xi_*\mu_j^2}\right)\right]^2\right)}{\prod_{i=1}^{p}\prod_{s=1}^{n}\exp\left(-\frac{\tau}{2}\left[y_{is}-\sum_{j=1}^{k}\frac{h_{js}}{(\xi_t\mu_j^2)^{\frac{1}{2}}}\exp\left(-\frac{(x_i-\mu_j)^2}{2\xi_t\mu_j^2}\right)\right]^2\right)}\right\}. \tag{3.5}$$

The move from ξ_t to ξ_* is accepted with this probability and the ξ parameter updated accordingly. Note that values of ξ outside the range (0, 0.01) have zero prior and are consequently rejected.

Updating μ

When updating a peak location μ the posterior changes in two places - both the likelihood and the prior contributions. Assume we are updating the j^{th} mean. Let $\mu_{j,(t)}$ be the j^{th} mean in the current set of parameters and $\mu_{j,*} \sim N(\mu_{j,(t)}, \sigma^2_{\mu_j})$ be the j^{th} mean in the proposed set of parameters. Thus the equation for the acceptance probability α in this case is:

$$\alpha(\lambda^{(t)}, \lambda^*) \quad =$$

$$\min\left\{1, \frac{e^{-\beta(N_R(\mu_{j,*}))}\prod_{i=1}^{p}\prod_{s=1}^{n}\exp\left(-\frac{\tau}{2}\left[y_{is}-\sum_{j=1}^{k}\frac{h_{js}}{(\xi\mu_{j,*}^2)^{\frac{1}{2}}}\exp\left(-\frac{(x_i-\mu_{j,*})^2}{2\xi\mu_{j,*}^2}\right)\right]^2\right)}{e^{-\beta(N_R(\mu_{j,(t)}))}\prod_{i=1}^{p}\prod_{s=1}^{n}\exp\left(-\frac{\tau}{2}\left[y_{is}-\sum_{j=1}^{k}\frac{h_{js}}{(\xi\mu_{j,(t)}^2)^{\frac{1}{2}}}\exp\left(-\frac{(x_i-\mu_{j,(t)})^2}{2\xi\mu_{j,(t)}^2}\right)\right]^2\right)}\right\} \tag{3.6}$$

where $N_R(\mu_{j,\cdot})$ is the number of peak locations within distance R of location $\mu_{j,\cdot}$. This step updates just one of the peaks and should be repeated for each location in the model.

Updating h

When updating a height scaling h only the likelihood changes. Similar to the updating procedure for ξ, the prior contribution remains the same since the prior on h is uniform. Let $h_{js}^{(t)}$ be the j^{th} scaling parameter of spectrum s in the current set of parameters and $h_{js}^{*} \sim N(h_{js}^{(t)}, \sigma^2_{h_{js}})$ be the j^{th} scaling parameter on spectrum s in the proposed set of parameters. Thus the equation for α is:

$$\alpha(\lambda^{(t)}, \lambda^{*}) \quad = \quad \min\left\{1, \frac{\prod_{i=1}^{p} \exp\left(-\frac{\tau}{2}\left[y_{is} - \sum_{j=1}^{k} \frac{h_{js}^{*}}{(\xi\mu_j^2)^{\frac{1}{2}}} \exp\left(-\frac{(x_i-\mu_j)^2}{2\xi\mu_j^2}\right)\right]^2\right)}{\prod_{i=1}^{p} \exp\left(-\frac{\tau}{2}\left[y_{is} - \sum_{j=1}^{k} \frac{h_{js}^{(t)}}{(\xi\mu_j^2)^{\frac{1}{2}}} \exp\left(-\frac{(x_i-\mu_j)^2}{2\xi\mu_j^2}\right)\right]^2\right)}\right\}, \tag{3.7}$$

and the proposed value of the js^{th} scaling h is accepted with this probability. As for the μ updates, this procedure only updates one peak height for one spectrum and should repeated for each h in turn in every MCMC iteration.

3.2.4 The Adapting Stage

All the proposal distributions used in MCMC algorithms in this chapter are $N(\mu, \sigma^2)$ where μ is the current value of the parameter. The value of σ^2 will determine how many of the proposed parameter values are accepted. The value of σ^2 can be initialised but it is difficult to know before starting the MCMC if this value is going to lead to the best acceptance rates.

To try to resolve this problem an *adapting* stage is built into the MCMC procedure. The proportion of acceptances for each parameter updated using Metropolis-Hastings should be around 40% to 60% to allow good mixing (Gelman, Roberts and Gilks, 1997). During the adapting stage the percentage of acceptances for each parameter is monitored and each of the proposal variances changed if the percentage is too low or too high. A general adapting procedure for a particular parameter is (Browne and Draper, 2000):

1. Run through 100 iterations of the MCMC algorithms and calculate the percentage of acceptances P_{acc} for that particular parameter in the last 100 iterations.

2. (a) if $P_{acc} < 50\%$ then decrease the variance of the proposal distribution to $\sigma^2_{new} = \sigma^2_{old}/(2 - (\frac{P_{acc}}{50}))$ and run for another 100 iterations.

   (b) if $P_{acc} > 50\%$ then increase the variance of the proposal distribution to $\sigma^2_{new} = \sigma^2_{old} \times (2 - (\frac{100-P_{acc}}{50}))$ and run for another 100 iterations.

3. When at least 300 iterations have been carried out (and every 100 iterations thereafter) check to see if the 3 most recent values of P_{acc} lie within the range $40\% < P_{acc} < 60\%$.

4. (a) if the 3 most recent values of P_{acc} are not all within the range then return to adapting the proposal variance as in step 2.

   (b) if they are then the variance of the proposal distribution is acceptable and is not changed further.

Then, once all the parameters have been initialised using this method, the actual burn in process and main iterations can take place. An upper bound is placed on the length of the adapting stage e.g. 5,000 iterations so that, in the rare occasion that the variances of the proposal distributions are not acceptable after such a time, the burn in period commences regardless.

3.2.5 Computational Speed Ups

The methods described in section 3.2.3 to update the parameter values are theoretically correct. However, when they are used exactly as described they are very computationally expensive. For example, when updating a height scaling parameter h, the calculation in equation (3.7) involves a complete recalculation of the posterior distribution with the proposed set of parameters. However, since we only update one parameter at a time most of the posterior distribution remains the same. The repeated calculation of all the other parts therefore just wastes time which could better be used running the algorithm for more iterations. Instead, to calculate the new posterior value we simply subtract the contribution which has changed and add in the new contribution.

When a peak is placed at a particular location the value of θ is calculated at every m/z value. When an m/z value is a large distance away from the peak location the value of θ is almost zero. The calculation of lots of things that are essentially zero also wastes time. To reduce the number of calculations needed we restrict how far a peak is allowed to have an influence. This restriction is set at a range of 6 of its standard deviations on either side of the mean. All other m/z values are then set to zero for that peak.

We could also speed up the time taken to model the data by using the High Performance Computing (GRID) system. Since each of the sections of data is analysed separately we could run them in parallel. There are many processors available and each section could be submitted to a different one. This would reduce the total time taken to the longest time taken for any one section. However, the parameters ξ and τ, which should be common across sections, will now have to modelled separately in each.

3.2.6 The Splitting Algorithm

Full evaluation of the likelihood as shown in equation (3.3) is very time consuming. In order to model the data within a reasonable time we wish to split the m/z values into distinct sections, where the dividing points are at low intensity values. We can then update the heights of peaks in one section without having to evaluate the whole likelihood.

Assume we have n spectra each of length p. The idea is to split this dataset into a partition of smaller sections of m/z values with lengths $p_1, p_2, \ldots, p_m$ such that $\sum_{i=1}^{m} p_i = p$.

It is possible to split the dataset in a huge number of possible places. However, many of these ways will split one peak over two sections. We should try to place split points so that the peaks in one section should not have much effect on other sections.

The main aim of the analysis is to identify where spectra in different groups have peaks of different heights. Biologically this could represent, for example, a molecule that is more abundant in cancer patients than non-cancer patients and which could be further analysed to introduce new drugs to treat the disease. However, there is only valuable information at and around peaks. There is no valuable information to be gained at an m/z value where the relative intensity is zero in all groups as this means that there are no molecules present at all. Hence a sensible place to suggest splitting the data is where the relative frequency of molecules at that m/z value is close to zero. To ensure that the split points have intensities as small as possible we consider the sum over all spectra. These sums are shown in figure 3.3 for the breast cancer and melanoma datasets respectively.

The splitting algorithm used is essentially binary - each section of the data is split into two parts by the next step. To ensure that the algorithm does not degenerate after the first split point is placed we need to be careful in how future points are chosen. If, for example, the first datapoint available has the lowest intensity then this would be chosen as the split point giving two parts, one with length 1. This is not desirable and so we limit the range of m/z values in which the split point can be found to the lower and upper quartiles of the m/z values. The complete algorithm is detailed below.

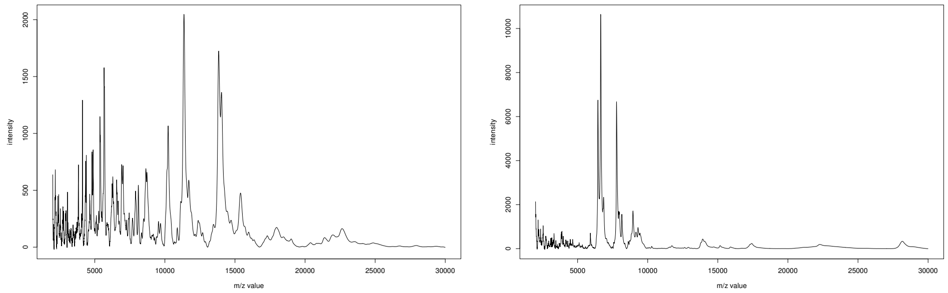

Figure 3.3: The sum of the data over all spectra in the breast cancer (left) and melanoma (right) datasets respectively.

1. Sum the data over all spectra at each m/z value.
2. The splitting step:
   - Calculate the lower quartile and upper quartile of the m/z values in each section of data.
   - Find the m/z value with the lowest intensity within each of these interquartile ranges. Place split points at these locations.
3. For each new section of data that contains more than 1,200 m/z values apply the splitting step again to this section.

The cut-off choice of 1,200 m/z values is arbitrary and was chosen as a tradeoff between the speed of fitting the model and keeping the number of sections relatively small. Using this algorithm the breast cancer data was split into 17 sections and the melanoma data into 19 sections. The split points chosen are shown in tables 3.1 and 3.2.

Each of the sections of data shown in tables 3.1 and 3.2 can now be analysed separately of the other sections in the dataset since we are assuming that the fitted model will be close to zero at the ends of each partition. However, this will involve the use of different τ and ξ parameters in each section.

m/z	3022	4187	5167	6056	7587	8266	9937	10728	13053
length	1115	1057	779	645	1008	414	952	423	1159

m/z	14948	16735	19128	19957	21110	24357	27149	29994
length	771	870	971	322	437	1170	945	914

Table 3.1: Split points obtained from the algorithm for the breast cancer data and the number of m/z values in each part.

m/z	2652	3621	4388	5668	7415	8497	9916	10692
length	737	942	659	982	1175	659	804	415

m/z	11367	13465	15652	16417	18604	19409	20695	23734
length	349	1023	985	328	898	317	493	1108

m/z	24790	26735	29994
length	368	658	1051

Table 3.2: Split points obtained from the algorithm for the melanoma data and the number of m/z values in each part.

3.3 Application to Datasets

The methods outlined in the previous two sections are now applied to both the breast cancer dataset and the melanoma dataset. Each dataset ranges over the same m/z values - 2,000 Da to 30,000 Da. The sections of data obtained by using the splitting algorithm detailed in the previous section were modelled separately and the results combined.

To find suitable starting values for the parameters we use the peak finding method detailed in chapter 2. This has two main benefits. Firstly if the parameters start in reasonable places the amount of time the algorithm takes to adapt the proposal variances may be reduced. Secondly, and more importantly, given starting values based on peak size, it is unlikely that a prominent peak will be ignored. With random starting values this may not be the case as parameters could converge elsewhere.

The number of peaks to be fitted to the data differed between each section. The number of visible peaks was counted and then this number was increased to account for any hidden peaks. This should not affect the results of the MCMC as the fitted heights of the extra peaks can be close to zero if no peak is present. For each section the adapting stage was used to fine-tune the parameter values before a burn in of 1,000 iterations and an MCMC run of 5,000 iterations. The adapting stage took 3,000 iterations on average before the proposal acceptance rates were suitable.

3.3.1 Results for the Breast Cancer Dataset

As explained in section 3.2, the model parameters included peak locations (μ_j), peak heights (h_{js}), a proportionality constant (ξ) used to model the peak variances, and a residual variance parameter ($1/\tau$). The current parameter values were checked at each iteration to determine the *maximum a posteriori* (MAP) parameter estimates.

In figure 3.4 one complete spectrum from the dataset is shown (black) along with the fitted model using the MAP parameter estimates (red). In total 138 peaks have been modelled. A smaller section of the dataset, between 7,600 Da and 8,300 Da, is shown in figure 3.5. This figure shows all the *adcon* spectra grouped by day. Similarly to figure 3.4, the black lines show the data and the red lines the MAP estimates, and table 3.3 shows the parameter values that are used to construct these MAP estimates.

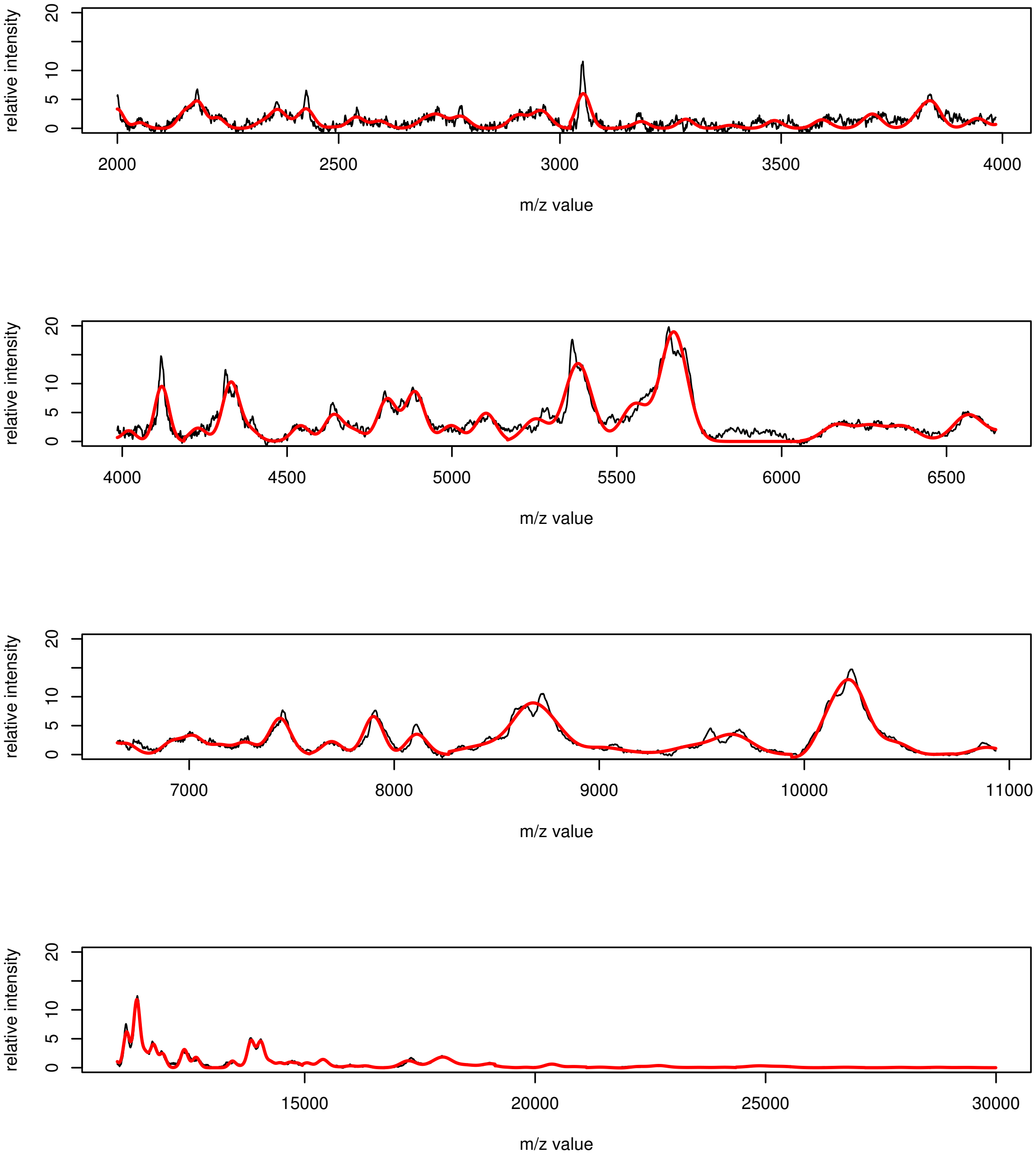

Figure 3.4: A single breast cancer spectrum and its fitted model. The black line shows a single spectrum from the breast cancer dataset and the red line shows the fitted model using the MAP parameter estimates.

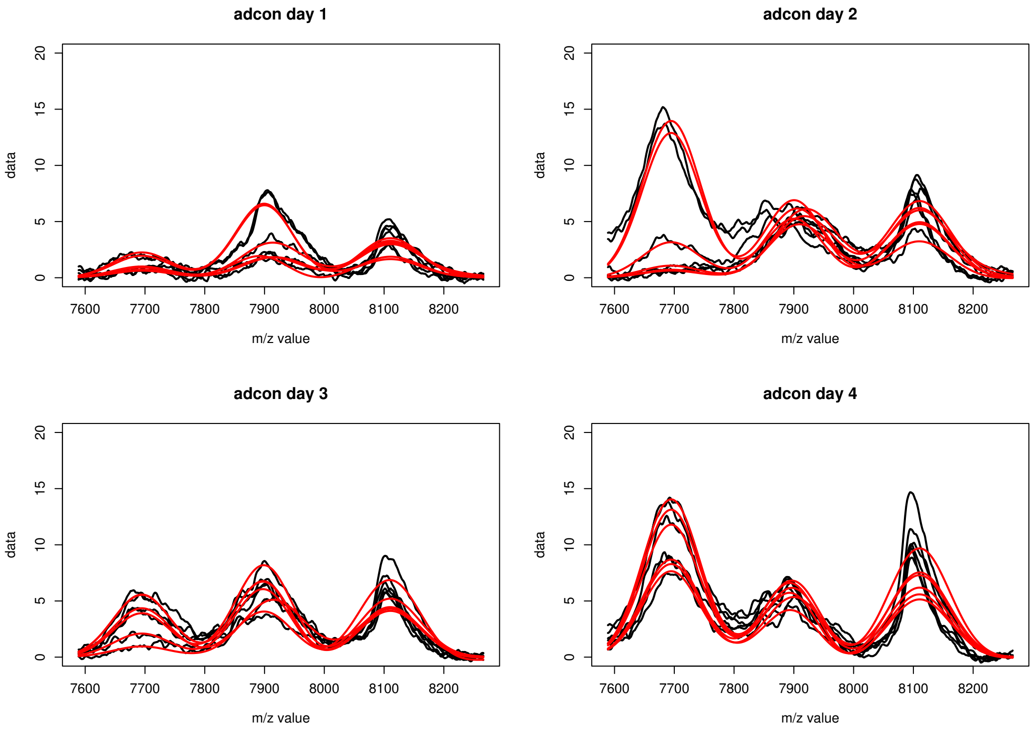

Figure 3.5: A section of the MAP after 5,000 iterations and original data for all 24 spectra in the *adcon* group on each day separately. The black lines show sections of the spectra from the breast cancer dataset and the red lines show the fitted models using the MAP parameter estimates.

parameter	ξ	τ	μ_1	μ_2	μ_3	μ_4	μ_5
MAP value	0.000039	3.5862	8110	7904	7694	7939	8212

Table 3.3: The MAP estimates of ξ, τ and the five location parameters μ_j for the section of the breast cancer data shown in figure 3.5.

The MAP estimates of θ shown in figures 3.4 and 3.5 seem to approximate the spectra quite well and the visible peaks have been identified by the model. However, some of the fitted heights are not very accurate for some peaks, for example the peak at around m/z value 8,100 Da on all four days. Here the modelled heights are not as large as the data. This could be attributed to the non-normal shape of this peak which is most evident on day 4.

Convergence of the parameters can be checked by inspecting trace plots. Figure 3.6 shows the trace plots for the precision parameter τ and the proportionality parameter ξ and figure 3.7 shows the trace plots for the five location parameters μ_j which relate to the section of data shown in figure 3.5.

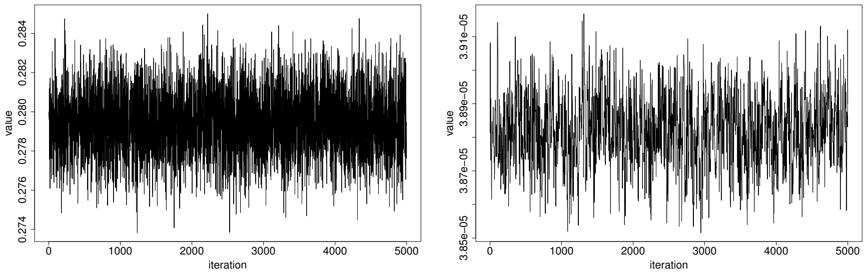

Figure 3.6: Trace plots of the precision parameter τ (left) and the proportionality constant ξ (right) for a section of the breast cancer dataset.

Trace plots for the 720 height scaling parameters are omitted. There appear to be no patterns in most of the traces which implies the chains are mixing well and that the adapting and burn-in period are of a suitable length. In the case of peak 5, shown in figure 3.7, the trace appears skewed in one direction. This is because this location was close to the end of this section of data and proposals out of the range were immediately rejected. The trace for peak 4 centres around an m/z value of 7,940 Da and appears to have a smaller acceptance probability than the others. This peak location is close to a more important peak in the model centred around 7,905 Da.

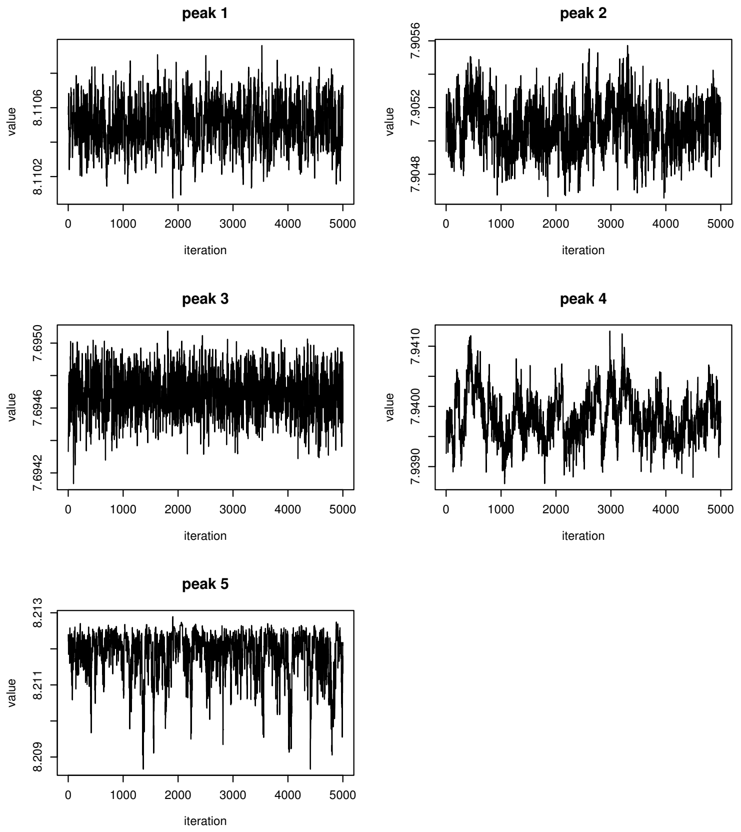

Figure 3.7: Trace plots of the five location parameters μ_j for a section of the breast cancer dataset.

To see how well the model fits the data the Akaike Information Criterion (AIC) statistic can be calculated (Akaike, 1974). The number of parameters in the model is comprised of 138 locations, 19,872 heights, 17 residual variances and 17 proportionality constants (each of the 17 sections had a separate estimate of ξ and τ). For the model used in this section the AIC is thus

$$\begin{aligned} -2 \quad \times & \quad \text{loglikelihood } + 2 \times \text{ no. of parameters} \\ &= (-2 \times 1,317,360) + (2 \times 20,044) \\ &= -2,594,632. \qquad (3.8) \end{aligned}$$

This value will be used in the rest of this chapter to determine whether more complex models are appropriate. Similarly, the BIC statistic for this model is $-2,343,817$. Comparisons of the models using the AIC and BIC statistics will be discussed later.

For each day and at each peak location in the model t-statistics (Student, 1908) were calculated using the maximum posterior estimates of the height parameters to identify which locations differed in height between the related pairs of control and treated cell-lines (e.g. *adcon* and *adtax*). Pairwise comparisons between the three control cell-lines (e.g. *tdcon* and *mccon*) were also calculated. Due to the large number of tests a false discovery rate algorithm was used (see section 1.6.6) to reduce the number of results identified as significant. The results of these analyses are shown in tables 3.4 and 3.5. The numbers in the tables indicate which day(s) had significant t-statistics after the FDR algorithm had been applied with $q^* = 0.05$. Similar statistics were also calculated which ignored the day information. However, under this assumption none of the locations differed significantly in height for any of the comparisons.

It appears from tables 3.4 and 3.5 that it is easier to distinguish between groups on days 3 and 4. This should be expected since the Taxol treatment will have had more time to take effect. In most cases of the control-treated comparisons the t-statistics were positive suggesting that the Taxol treated cell-lines have smaller heights. Also it should be noted that significant differences between *adcon* and *adtax* cell-lines do not occur very often - the only exception occurring at an m/z value of 5,407 Da. This should also not be surprising since the MCF-7/ADR cell-lines are meant to be resistant to Taxol.

We should be careful when using these m/z values to show differences between groups as inference is based on only 6 observations per group. Tests using such small amounts of data lack power and much more data would be needed to obtain reliable biomarkers.

m/z	2959	3053	3709	3809	3835	4019
adc/tdc				(3)		(4)
adc/mcc		(3)				(2)
tdc/mcc	(1)		(4)	(4)	(3,4)	(3,4)

m/z	4119	4388	4541	4641	4703	4803
adc/tdc		(4)			(4)	
adc/mcc	(4)	(3,4)		(2,3)		(2,3,4)
tdc/mcc		(3,4)	(4)	(3,4)	(2,4)	(4)

m/z	4887	5103	5253	5376	5669	6916
adc/tdc		(4)	(3)		(4)	
adc/mcc						
tdc/mcc	(1,2,3,4)		(2,4)	(3,4)	(2)	(3,4)

m/z	7147	7443	7694	7939	8110	8212
adc/tdc		(3)	(4)	(4)		(4)
adc/mcc				(4)	(2,4)	(4)
tdc/mcc	(2,4)	(2,3)	(3,4)			

m/z	9187	10016	10230	10435	11137	11357
adc/tdc						(3)
adc/mcc	(3)				(4)	
tdc/mcc	(3,4)	(2)	(2,3)	(3)		(3)

m/z	13636	14046	14250	15402	17252	17764
adc/tdc		(4)	(4)			(4)
adc/mcc						
tdc/mcc	(3)	(3)	(3)	(3,4)	(3)	(2,4)

m/z	19030	20360	21010	23083
adc/tdc	(2)	(4)	(4)	
adc/mcc				
tdc/mcc	(1)			(4)

Table 3.4: The peak locations with significant t-statistics between the three pairs of control groups.

m/z	3592	4389	4703	5407	5669	6369
adc/adt				(3)		
tdc/tdt	(2)	(4)	(4)		(2)	
mcc/mct				(4)		(4)

m/z	7019	8110	8473	10228	11719	12389
adc/adt						
tdc/tdt			(4)	(3)	(3)	
mcc/mct	(4)	(4)				(4)

m/z	12643	13432	13636	14046	14460	14856
adc/adt						
tdc/tdt				(3)	(3)	
mcc/mct	(4)	(4)	(4)	(4)		(4)

m/z	15043	15402	16315	17983	18506	19030
adc/adt						
tdc/tdt	(3)			(1)		
mcc/mct	(4)	(4)	(4)		(4)	(4)

m/z	20815	21010	22227	23083	24117	25373
adc/adt						
tdc/tdt		(3)	(1)	(3)		
mcc/mct	(4)			(4)	(4)	(4)

Table 3.5: The peak locations with significant t-statistics between the pairs of control and treated cell-lines.

3.3.2 Results for the Melanoma Dataset

One complete melanoma spectrum (black) is shown in figure 3.8 along with the fitted model using the MAP parameter estimates (red). In total 112 peaks have been modelled. To see the model more clearly, a smaller section of the dataset, between 7,400 Da and 8,500 Da, is shown in figure 3.9.

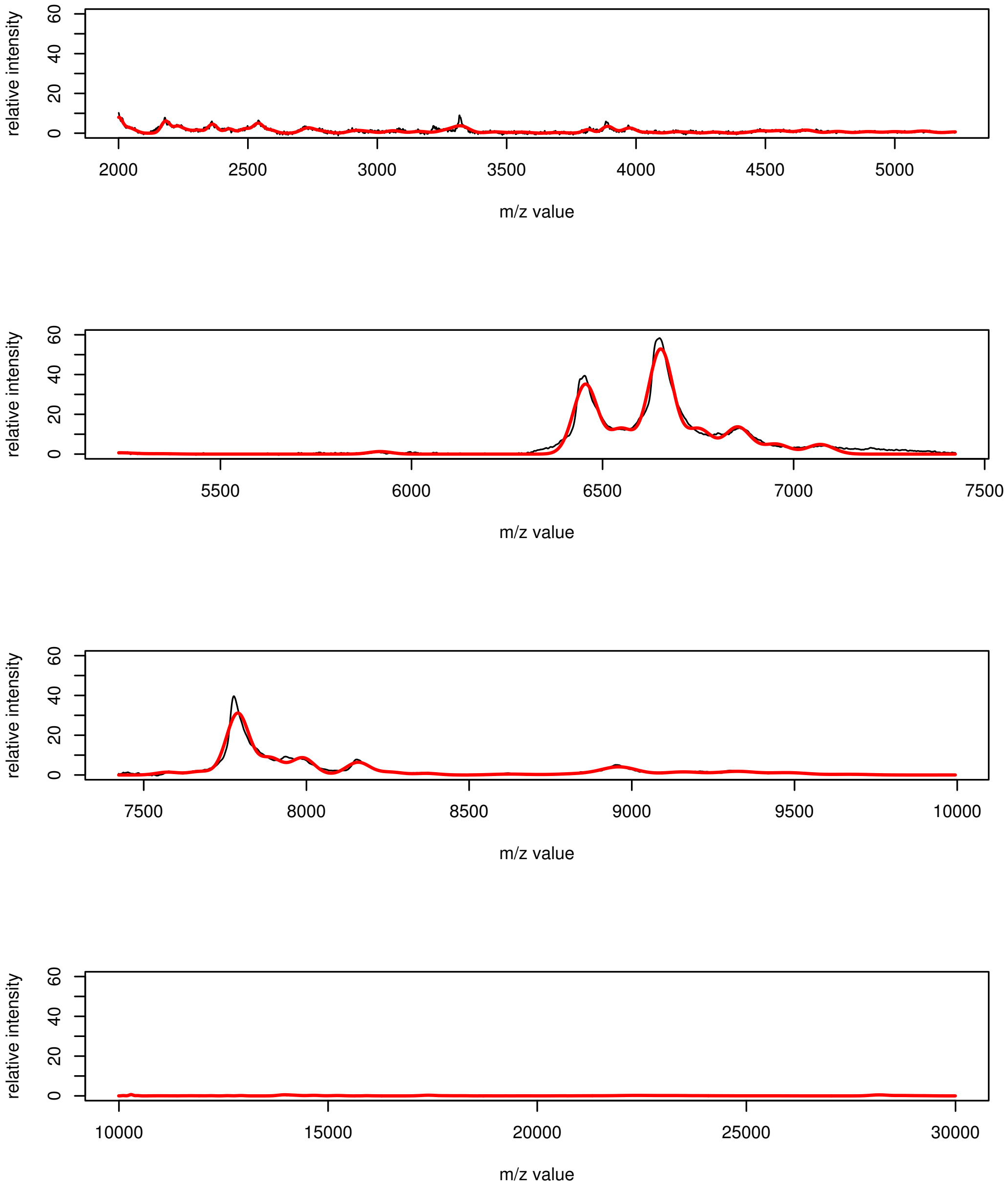

Figure 3.8: A single melanoma spectrum and its fitted model. The black line shows a single spectrum from the melanoma dataset and the red line shows the fitted model using the MAP parameter estimates.

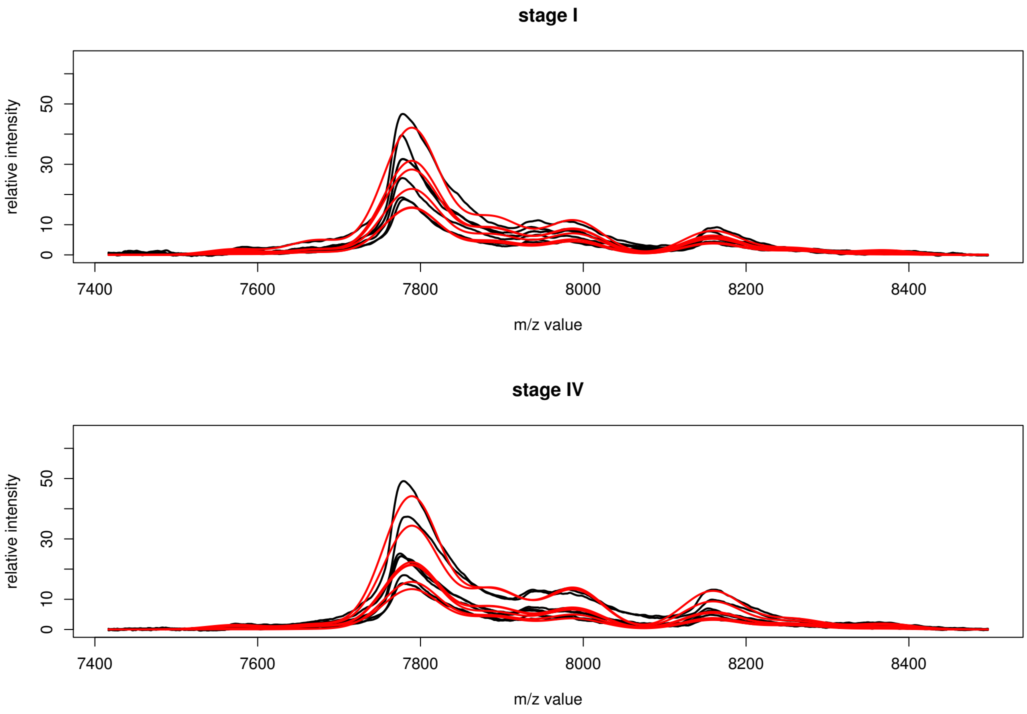

Figure 3.9: A section of the MAP after 5,000 iterations and original data for 6 spectra in each of the melanoma stages separately.

parameter	ξ	τ
MAP value	0.000020	0.4306

parameter	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8
MAP value	7788	7988	8158	7888	7672	8368	8258	7572

Table 3.6: The MAP estimates of ξ, τ and the eight location parameters μ_j for the section of the melanoma data shown in figure 3.9.

Figure 3.9 shows 6 spectra from each of the stages of melanoma. Again, the black lines show the data and the red lines the MAP estimates. The MAP estimates for each of the model parameters are given in table 3.6.

The MAP estimates of θ_{is} model the data well albeit with a similar problem to the breast cancer data in that some of the fitted heights are not sufficiently large to correctly match the data. This can be seen in figure 3.9 where the MAP at an m/z value around 7,800 Da falls short of the datapoints.

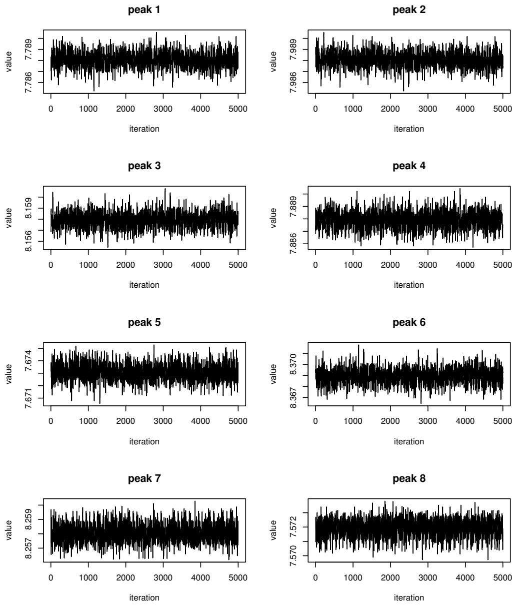

Figure 3.10: Trace plots of the eight location parameters μ_j for the melanoma dataset.

We again check convergence of the parameters by inspecting trace plots. Figure 3.10 shows the trace plots for the eight location parameters μ_j and figure 3.11 shows the trace plots for the precision parameter τ and the proportionality parameter ξ. Trace plots for the 1,640 height scaling parameters are omitted. In general the traces are acceptable as there are no obvious patterns.

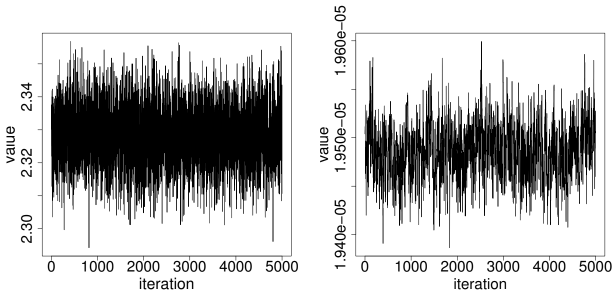

Figure 3.11: Trace plots of the precision parameter τ (left) and the proportionality constant ξ (right) for a section of the melanoma dataset.

In table 3.7 are the m/z values with significant t-statistics between the two stages of melanoma after the FDR algorithm has been applied with $q^* = 0.05$. The majority of the statistics are positive which suggests that as the stage of melanoma increases the abundance of molecules at these m/z values decreases. All of the significant peaks in the range 11,300 - 12,200 Da have negative t-statistics which suggests an increase in the number of molecules between stage I and stage IV. The melanoma results are more reliable than the breast cancer results as they are based on a larger amount of information.

The AIC statistic can be calculated as for the breast cancer dataset. The number of parameters in the model is comprised of 112 locations, 22,960 heights, 19 residual variances and 19 proportionality constants (each of the 19 sections had a separate estimate of ξ and τ) and thus for the single peaks model the AIC is $(-2 \times 2,811,217) + (2 \times 23,110) = -5,576,214$. Similarly, the BIC statistic for this model is $-5,278,873$. Comparisons of the models using the AIC and BIC statistics will be discussed later.

In the fitted models for both the breast cancer and the melanoma there are some height parameters which are not as large as the data. To try to improve the fit more complex models are now introduced.

m/z	2227	2262	2306	2488	2539	2766
t-statistic	8.077	6.316	5.400	7.885	5.741	5.687

m/z	2962	3298	3556	3828	3888	3974
t-statistic	7.310	5.231	4.681	8.500	9.085	4.294

m/z	4478	4656	4778	5107	6455	6652
t-statistic	4.198	4.836	4.898	4.738	5.860	5.342

m/z	6752	7572	7672	7788	7988	8158
t-statistic	-5.408	7.988	7.443	7.410	5.769	5.500

m/z	8368	8928	9323	9481	9670	11528
t-statistic	5.043	4.845	4.860	5.458	4.555	-5.499

m/z	11704	13946	14210	14659	17234	17424
t-statistic	-5.471	5.155	6.951	6.350	4.619	5.151

m/z	17810	18207	28172	28745	29233
t-statistic	5.270	4.525	7.566	6.905	6.479

Table 3.7: The peak locations with significant t-statistics between the two stages of melanoma.

3.4 Double Peaks

As pointed out in section 3.3.1, the peak present in the breast cancer dataset at an m/z value of 8,100 Da in figure 3.5 does not look Gaussian and thus the model does not fit well to the data at this location. On further examination of this peak it appears to possibly consist of a combination of two peaks at the same location. The first is a peak with small variance which accounts for the spikiness of the overall peak, and the second is a peak with larger variance which contributes a baseline amount of intensity.

In the model discussed in section 3.2.1 peaks were fitted at certain m/z values and the standard deviation of these peaks was proportional to the location. To take the combination peaks into account a slight alteration of the model is needed. At each peak location two peaks will now be fitted. The first will be fitted as in the previous model and the second will be fitted to exactly the same location but will have an increased variance.

3.4.1 The Double Peaks Model

Each datapoint y_{is} is distributed as $y_{is} \sim N(\theta_{is}, \tau^{-1})$ where θ_{is} is changed to:

$$\begin{aligned}\theta_{is} &= \sum_{j=1}^{k} h_{js}(\xi\mu_j^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\xi\mu_j^2)^{-1}(x_i-\mu_j)^2\right) \\ &\quad + \sum_{j=1}^{k} h^*_{js}(\omega\xi\mu_j^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\omega\xi\mu_j^2)^{-1}(x_i-\mu_j)^2\right) \qquad (3.9)\end{aligned}$$

where the index $i = 1, \ldots, p$ represents the position on the spectrum, x_i is the i^{th} m/z value, $s = 1, \ldots, n$ is the spectrum number, $j = 1, \ldots, k$ is the peak number, ξ is the constant of proportionality and μ_j are the peak locations. Note that the two peaks have the same location but the second peak has a larger variance by a factor of $\omega > 1$. From preliminary analysis it appeared that $\omega = 2$ was a reasonable choice. We fix this parameter initially and examine different variances in section 3.5.

The parameters h_{js} represent the scaling that adjusts the height of the peaks with small variance and the parameters h^*_{js} are similar for the peaks with large variance.

As in the previous section, the model will be fitted to each section of data separately and the results pooled to obtain an overall model. The results are presented in the next section.

3.4.2 Results for the Breast Cancer Dataset

To compare with the single peaks model 138 peaks have again been modelled. A small section of the dataset, between 7,600 Da and 8,300 Da, is shown in figure 3.12. This figure shows one spectrum in the *adcon* group from day 4. The black line shows the data, the red line shows the MAP estimate of θ under the single peaks model and the green line shows the MAP estimate under the double peaks model. Table 3.8 shows the parameter values that give the MAP estimate under the double peaks model.

Comparing the two models we see that the double peaks model provides a better fit to the data than the single peaks model for the peak at 8,100 Da as the fitted heights are increased, better matching the data. However, the fit is still not perfect as the fitted heights remain smaller than the data. When the MAP parameter estimates in table 3.8 are compared with those for the single peaks model it can be seen that the peak locations have not changed greatly. The proportionality parameter has decreased from its value in the single peaks model. Since ξ no longer has to solely account for the variance of the complete peaks a smaller value allows the peaks with smaller variance to be modelled more closely.

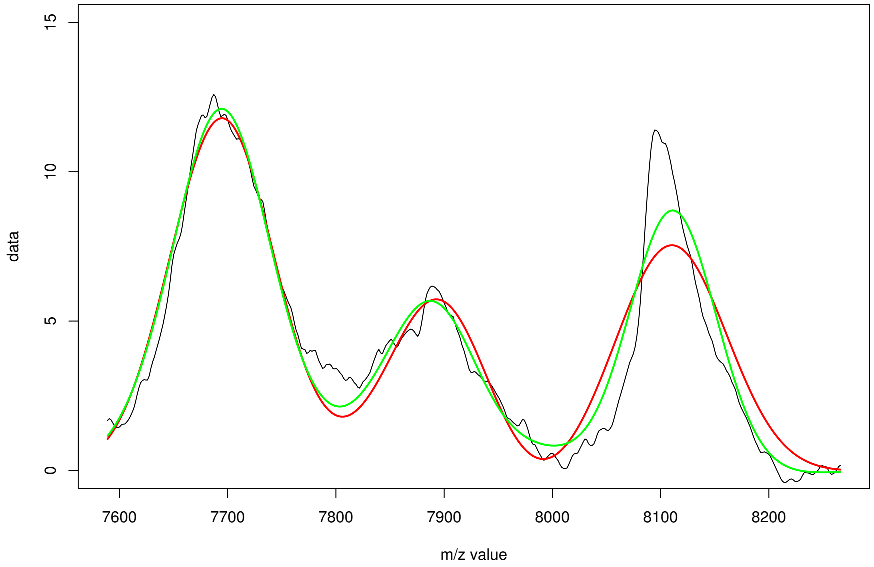

Figure 3.12: A section of the MAP after 5,000 iterations using single peaks (red), double peaks (green) and original data for one spectrum in the *adcon* group from day 4.

parameter	ξ	τ	μ_1	μ_2	μ_3	μ_4	μ_5
MAP value	0.000024	5.5413	8111	7911	7694	7936	8211

Table 3.8: The MAP parameter estimates for the section of the breast cancer data shown in figure 3.12 using the double peaks model (green).

Trace plots for all the modelled parameters were checked. There were no evident patterns except for the one sided proposals for one peak due to its proximity to the end of the section. To compare this model with the one in the previous section we recalculate the AIC. The number of parameters in the model is comprised of 138 locations, 39,744 heights, 17 residual variances and 17 constants. For the double peaks model the AIC is thus $(-2 \times 1,585,413) + (2 \times 39,916) = -3,090,994$. This value is lower than that of the single peaks model and so the double peaks model is an improvement when using this criterion. The BIC statistic for this model is $-2,591,523$.

As with the single peaks model t-statistics were calculated between pairs of control groups (on each day separately and combined) and also between similar control/treated pairs (on each day separately and combined). The tests were carried out on the MAP estimates of θ at each peak location and the false discovery rate algorithm was again used because of multiple testing. The peaks exhibiting significant differences between the groups remain almost identical to the ones identified in tables 3.4 and 3.5. Again combining days resulted in no significant results.

3.4.3 Results for the Melanoma Dataset

Figure 3.13 shows one spectra from the melanoma dataset over the same range of m/z values as examined previously. The black line shows the data and the red and green lines the MAP estimates of θ under the single and double peaks models respectively. The MAP parameter estimates under the double peaks model which were used to construct the green curve are shown in table 3.9.

From figure 3.13 we can see that the double peaks model again provides a better fit to the data than the single peaks model for the peak at 7,800 Da as the fitted heights are increased. The MAP estimate of θ for the area between m/z values of 7,900 Da and 8,000 Da is more accurate.

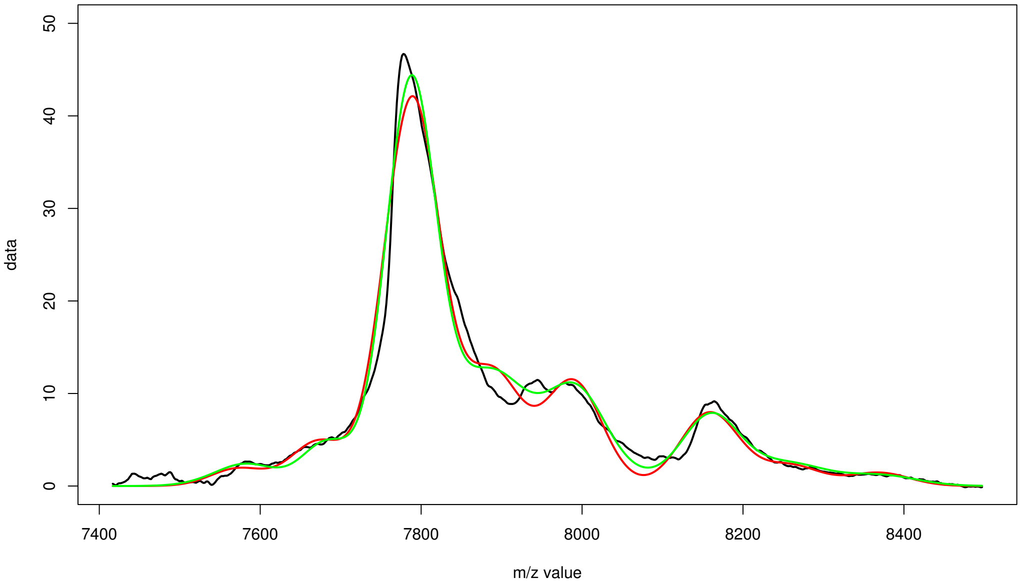

Figure 3.13: A section of the MAP after 5,000 iterations using single peaks (red), double peaks (green) and original data for one spectrum in the melanoma dataset.

parameter	ξ	τ
MAP value	0.000015	0.8547

parameter	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8
MAP value	7787	7887	8160	7991	7681	8370	8260	7581

Table 3.9: The MAP parameter estimates for the section of the melanoma data shown in figure 3.13 using the double peaks model (green).

No patterns were visible in the trace plots for the parameters so the chains were mixing in an acceptable manner.

To find peaks with significantly different heights between the two stages of melanoma t-statistics were calculated using the MAP estimates of θ at each peak location and the FDR algorithm. The significant locations remain essentially the same as in the single peaks model and the signs of the majority of significant t-statistics remain positive showing that the heights of the peaks in stage IV are lower than in stage I.

To compare with the previous model for this dataset we calculate the AIC. The model contains 112 location parameters, 45,920 heights, 19 residual variances and 19 constants. For the double peaks model the AIC is thus $(-2 \times 2,908,745) + (2 \times 46,070) = -5,725,350$. Similarly to the results for the breast cancer data this AIC value is much lower than for the single peaks model and we conclude that, out of these two options, the double peaks model is much more preferable using this criterion. The BIC statistic for the double peaks model is $-5,132,598$. The statistic has increased compared with that in the single peaks model as a penalty has been imposed for the introduction of nearly twice the number of parameters.

A more complex model will now be developed which incorporates the non-symmetry of some of the peaks.

3.5 Peak Offsetting

When analysing a substance by mass spectrometry it is common practice to carry out two sets of analyses - with different resolutions. The resolution will determine how well the mass spectrometer can differentiate between molecules at m/z values close to each other. The first analysis is carried out with a standard resolution and the results studied to find sections of the spectrum that warrant further attention. The second analysis is then carried out with a much higher resolution on only these sections of data. More precise data is obtained and more accurate differences can be identified. See figure 3.14 for an example of a chemical compound's mass spectrum showing a single peak studied at higher resolution.

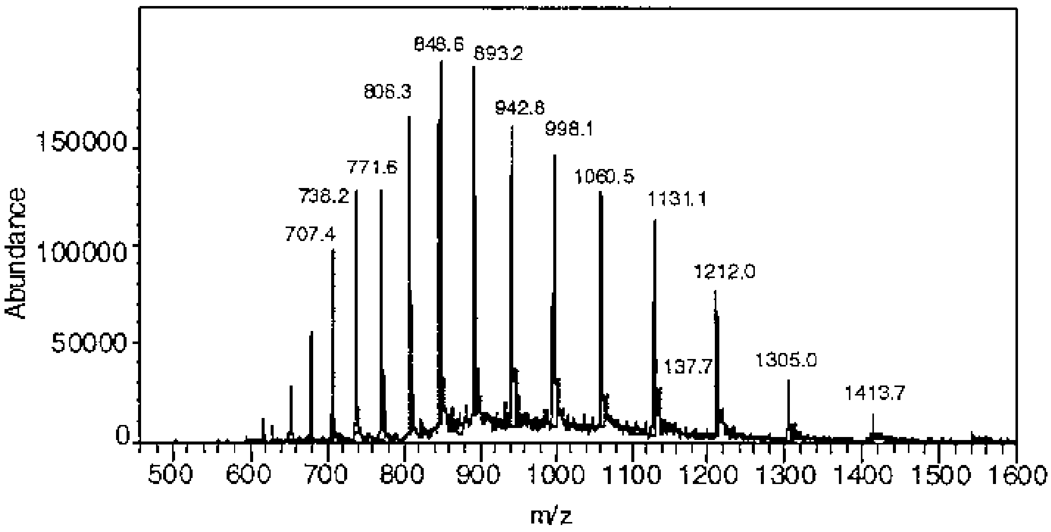

Figure 3.14: The mass spectrometry results for a single peak using a machine with higher resolution. From Wiley, J. & Sons (2006)

The data we are considering was obtained using a mass spectrometer with a standard resolution of 0.2%. This results in spectra resembling the 'series of Gaussian peaks' that has been mentioned already. As shown in figure 3.14 a peak may consist of many *spikes*. Spikes within the same peak can represent either a completely different molecule or the same overall molecule with slight modifications.

An *m/z* value is essentially a measure of mass. The mass of a molecule can be changed in two ways. Firstly an atom can have larger mass than normal if it is an isotopic variant. For example, there are 3 different isotopes of hydrogen - hydrogen, deuterium and tritium - which have relative masses of 1, 2 and 3 respectively. If a molecule contains some of these isotopic variants it will have greater mass and thus its spike will appear at a different *m/z* value to the original molecule. Secondly post-translational modifications can occur. Before mass spectrometry of the sample takes place it is possible that the structure of some proteins could change, for example via the attachment of another functional group. When the results are obtained the spike for the new bigger ion will appear at an *m/z* value greater than the original.

From this it can be seen that fitting a double peaks model is a move in the right direction since many spikes make up the overall peak. However, the restriction of equating locations is not necessarily sound. The peaks in the data are much more likely to have longer right hand tails due to isotopic variation and post-translational modifications. To account for this the locations of the peaks with larger variance could be altered from μ_j to $\mu_j + \delta_j$. The double peaks model is adjusted accordingly in the next section.

3.5.1 The Offset Peaks Model

To model the offset peaks we slightly alter the equation for θ_{is} described in section 3.2.1. Each peak location still combines two peaks - the first will be fitted as in both the previous models and the second will be fitted to a location within δ_j to the right of the location μ_j.

Each datapoint y_{is} is distributed as $y_{is} \sim N(\theta_{is}, \tau^{-1})$ where θ_{is} is now:

$$\begin{aligned}\theta_{is} &= \sum_{j=1}^{k} h_{js}(\xi\mu_j^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\xi\mu_j^2)^{-1}(x_i-\mu_j)^2\right) \\ &\quad + \sum_{j=1}^{k} h^*_{js}(\omega\xi(\mu_j+\delta_j)^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\omega\xi(\mu_j+\delta_j)^2)^{-1}(x_i-(\mu_j+\delta_j))^2\right)\end{aligned} \tag{3.10}$$

where the index $i = 1, \ldots, p$ represents the position on the spectrum, x_i is the i^{th} m/z value, $s = 1, \ldots, n$ is the spectrum number, $j = 1, \ldots, k$ is the peak number, ξ is the constant of proportionality and μ_j are the peak locations. The δ_j parameters model the offset from the main location μ_j of the second peaks. It is possible for a δ_j to be zero if no offset is present. The variance scaling factor for the double peaks is modelled by ω. The parameters h_{js} represent the scaling that adjusts the height of the peaks with small variance and the parameters h^*_{js} are similar for the offset peaks with larger variance.

The MCMC algorithm is altered to incorporate another Metropolis-Hastings step for the offset parameters δ_j. A uniform $\delta_j \sim Unif(0, 0.1)$ prior was used for these parameters. Since we are only considering offsets to the right the minimum offset possible is 0. The maximum offset is set to be 0.1 since it is possible for another peak to be present if the m/z value is greater than 100 Da away. When updating a δ_j the only part of the posterior that actually changes is the contribution from the likelihood. The prior contribution remains the same since the prior on δ_j is uniform. Assume that the current value of δ_j is $\delta_{j,t}$ and the proposed value is $\delta_{j,t+1} \sim N(\delta_{j,t}, \sigma^2_\delta)$. Then the equation for α, the acceptance probability, is:

$$\alpha(\lambda^{(t)}, \lambda^*) \quad = \quad min \left\{ 1, \frac{\prod_{i=1}^{p} \prod_{s=1}^{n} \exp\left(-\frac{\tau}{2}\left[y_{is} - \theta^*_{is}\right]^2\right)}{\prod_{i=1}^{p} \prod_{s=1}^{n} \exp\left(-\frac{\tau}{2}\left[y_{is} - \theta^{(t)}_{is}\right]^2\right)} \right\}. \tag{3.11}$$

where θ^*_{is} is the proposed value of θ_{is} (see equation (3.10)) under the proposed set of parameter values. The move from δ_t to δ_* is accepted with this probability and the δ_j parameter updated accordingly. Note that values of δ outside the range (0, 0.1) are assumed to have zero prior and are not considered. This step is repeated for each of the peaks.

3.5.2 Offset Peaks and Differing Variances

In the following results section we consider two possibilities. Firstly we maintain the restriction that the ω parameters are all equal to 2 and secondly, to make the model as general as possible, we allow these variance scaling parameters to vary at each peak. For this new model the ω parameters will have a prior distribution $\omega \sim Unif(1.5, 10)$. We do not wish the variances of the two peaks at the same location to become equal so the lower bound was set to be 1.5. The upper bound was set to 10 to allow locations to mainly consist of a single peak.

The MCMC algorithm is altered to incorporate another Metropolis-Hastings step for the variance scaling parameters ω_j. When updating an ω_j the posterior only changes through the likelihood. The prior contribution remains the same since the prior on ω_j is uniform. Assume that the current value of ω_j is $\omega_{j,t}$ and the proposed value is $\omega_{j,t+1} \sim N(\omega_{j,t}, \sigma_\omega^2)$. Then the acceptance probability α is calculated as in equation (3.11) and the move from $\omega_{j,t}$ to $\omega_{j,t+1}$ is accepted with this probability. Note that values of ω_j outside the range (1.5,10) have zero prior and are not considered. This step is repeated for each of the peaks.

In summary, the four models considered in this chapter are:

	model	section
I	single Gaussian peaks	3.2
II	double peaks mixture model	3.4.1
III	offset peaks mixture model	3.5.1
IV	offset peaks with differing variances	3.5.2

Table 3.10: A summary of the four models considered in this chapter.

The results from these models will now be presented.

3.5.3 Results for the Breast Cancer Dataset

To enable suitable comparisons to be drawn with the previous models, 138 peaks have again been modelled. Figure 3.15 shows the same section of data as before but with the MAP estimates of θ under all four models considered in this chapter. The black lines show the data, the red lines the MAP estimates of θ under the single peaks model, the green line the MAP under the double peaks model, the turquoise line the MAP under the offset peaks model and the blue line the MAP under the differing variances model. Table 3.11 shows the parameter values that give the MAP estimate under the offset peaks model (turquoise) and table 3.12 the parameter values under the differing variance model (blue).

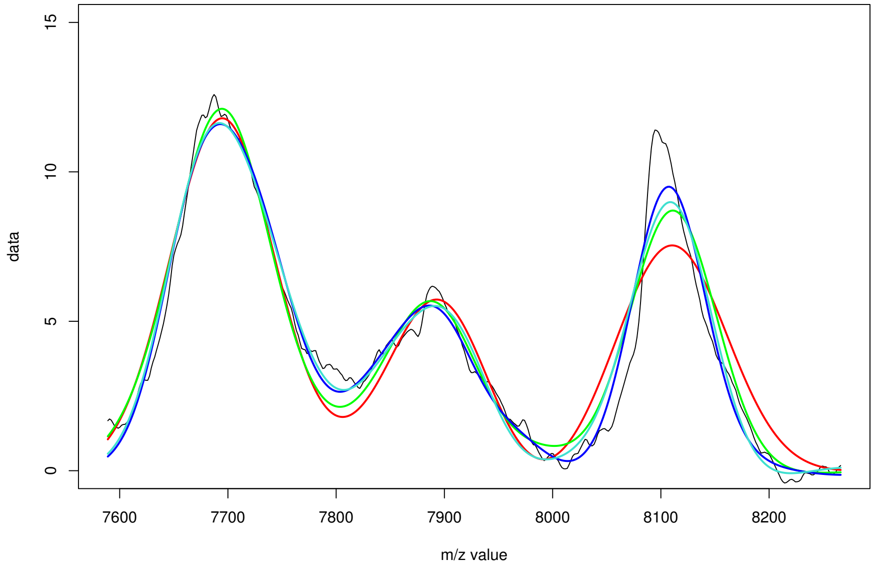

Figure 3.15: A section of the MAP after 5,000 iterations using all the models and original data for one spectrum in the *adcon* group for day 4. The red line shows the MAP under the single peaks model, the green line the double peaks model, the turquoise line the offset peaks model and the blue line the differing variances model.

parameter	ξ	τ
MAP value	0.000022	8.671

parameter	μ_1	μ_2	μ_3	μ_4	μ_5
MAP value	8106	7834	7664	7931	8265

parameter	$\mu_1 + \delta_1$	$\mu_2 + \delta_2$	$\mu_3 + \delta_3$	$\mu_4 + \delta_4$	$\mu_5 + \delta_5$
MAP value	8141	7920	7711	7962	8266

Table 3.11: The MAP parameter estimates for the section of the breast cancer data shown in figure 3.15 under the offset peaks model (turquoise).

parameter	ξ	τ
MAP value	0.000017	9.703

parameter	μ_1	μ_2	μ_3	μ_4	μ_5
MAP value	8110	7843	7671	7951	8264

parameter	$\mu_1 + \delta_1$	$\mu_2 + \delta_2$	$\mu_3 + \delta_3$	$\mu_4 + \delta_4$	$\mu_5 + \delta_5$
MAP value	8137	7922	7720	8151	8265

parameter	ω_1	ω_2	ω_3	ω_4	ω_5
MAP value	1.50	1.74	2.03	3.42	1.68

Table 3.12: The MAP parameter estimates for the section of the breast cancer data shown in figure 3.15 under the differing variances model (blue).

From figure 3.15 we can see that the offset peaks model again provides a better fit to the data than the previous models for the peak at 8,100 Da. The fitted heights are increased further, better matching the data although there still remains some difference. For the purposes of model comparison we recalculate the AIC. The number of parameters in the model is comprised of 138 locations, 39,744 heights, 138 offset parameters, 17 residual variances and 17 proportionality constants. For the model used in this section the AIC is thus $(-2 \times 2,182,023) + (2 \times 40,056) = -4,283,938$. This value is lower than that of the double peaks model and so the offset peaks model is a further improvement. The BIC statistic for the offset peaks model is $-3,782,705$. The statistic has again lowered from that in the double peaks model and so the offset peaks model is an improvement.

The difference between the data and the fitted model is slightly reduced by the use of the more complicated model with differing variances, however, this new model does not appear to give a much better fit to the data shown than that under the offset model. The more complicated model must fit better in other places, however, as we see a reduction in the AIC and BIC. The number of parameters in the model is comprised of 138 locations, 39,744 heights, 138 offset parameters, 138 variance scaling parameters, 17 residual variances and 17 proportionality constants. For the model used in this section the AIC is thus $(-2\times 2,188,279)+(2\times 40,192) = -4,296,174$ and the BIC is $-3,793,243$.

Trace plots for all the modelled parameters were checked in both models. There were no evident patterns except for the proposals only accepted in one direction for one peak due to its proximity to the end of the section.

When the MAP parameter estimates in tables 3.11 and 3.12 are compared with those for the previous models we see that the peak locations have again not changed much. The offset parameters for visible peaks are all positive and the variance scaling parameters are all around the value 2. For the fifth peak it is seen that the value of δ_5 is zero as the offset peak is fitted to effectively the same location. The value of the proportionality parameter has decreased although not by as much as last time.

The t-statistics using the MAP estimates of θ between pairs of control groups and between related control/treated groups did not show any significantly different peak locations compared with those of the original single peaks model for either of the offset models, although the absolute value of a large number of the t-statistics has increased. Comparisons which ignored day information all remained insignificant. The significant peak locations after correcting for multiple testing can be seen in tables 3.4 and 3.5.

3.5.4 Results for the Melanoma Dataset

The same small section of the dataset as examined previously is shown in figure 3.16 along with the MAP estimates of θ for each of the four models. Figure 3.16 shows one spectrum from the melanoma dataset - the black line shows the data and the red, green, turquoise and blue lines the MAP estimates of θ under the single peaks, double peaks, offset peaks and differing variance peaks models respectively. Tables 3.13 and 3.14 show the parameter values that are used to construct the MAP estimates of θ under the offset peaks model (turquoise) and differing variances model (blue).

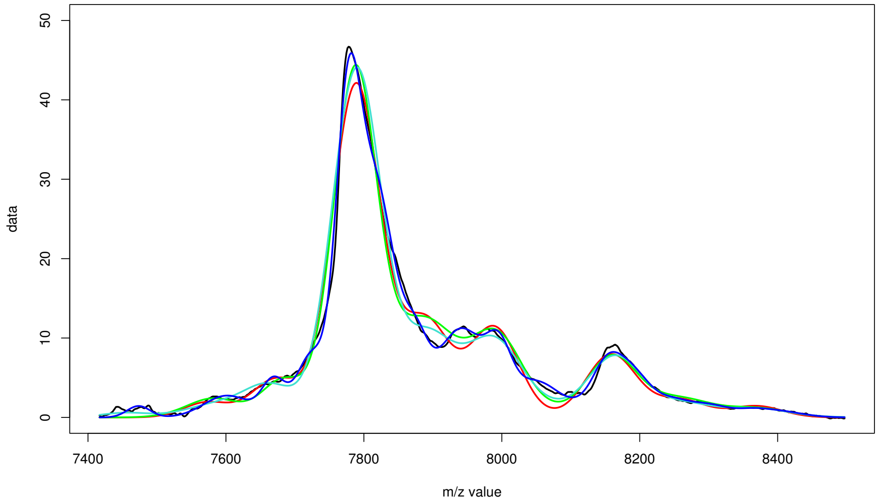

Figure 3.16: A section of the MAP after 5,000 iterations using offset peaks and original data for one spectra in the melanoma dataset. The red line shows the MAP under the single peaks model, the green line the double peaks model, the turquoise line the offset peaks model and the blue line the differing variances model.

parameter	ξ	τ
MAP value	0.000012	1.526

parameter	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8
MAP value	7789	8163	8267	7989	7665	8369	7889	7453

parameter	$\mu_1+\delta_1$	$\mu_2+\delta_2$	$\mu_3+\delta_3$	$\mu_4+\delta_4$	$\mu_5+\delta_5$	$\mu_6+\delta_6$	$\mu_7+\delta_7$	$\mu_8+\delta_8$
MAP value	7869	8196	8267	8035	7748	8372	7954	7544

Table 3.13: The MAP parameter estimates for the section of the melanoma data shown in figure 3.16 under the offset peaks model (turquoise).

parameter	ξ	τ
MAP value	0.000005	2.422

parameter	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8
MAP value	7775	7875	8149	7996	7669	8363	8249	7473

parameter	$\mu_1+\delta_1$	$\mu_2+\delta_2$	$\mu_3+\delta_3$	$\mu_4+\delta_4$	$\mu_5+\delta_5$	$\mu_6+\delta_6$	$\mu_7+\delta_7$	$\mu_8+\delta_8$
MAP value	7814	7940	8175	8048	7723	8392	8290	7502

parameter	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
MAP value	2.04	2.69	3.40	3.06	1.60	1.89	2.71	2.39

Table 3.14: The MAP parameter estimates for the section of the melanoma data shown in figure 3.16 under the differing variances model (blue).

From figure 3.16 we can see that the offset peaks model does not appear to show much difference in the range shown compared with the double peaks model. However, when considering the AIC we see a reduction so the model provides a better fit in some areas of the data using this criterion. The number of parameters is 46,182 (112 locations, 45,920 heights, 112 offset parameters, 19 residual variances and 19 proportionality constants). For the model used in this section the AIC is thus $(-2 \times 3,136,830) + (2 \times 46,182) = -6,181,296$. This lower value indicates that the offset peaks model is an improvement upon the double peaks model. The BIC statistic is $-5,587,103$ which is lower than that of both the single and double peaks models.

When comparing the MAP estimates of θ from the differing variances model with those from the simpler offset peaks model it can be seen that the more complicated model gives a much better fit to the data shown. The height of the peak at around 7,800 Da is much more closely fitted and the range between 7,900 and 8,000 Da is an improvement over all the previous models. This is borne out by a reduction in the AIC. The number of parameters in the differing variances model is 46,294 (112 locations, 45,920 heights, 112 offset parameters, 112 variance scaling parameters, 19 residual variances and 19 proportionality constants). For this model the AIC is therefore $(-2\times 3,146,468)+(2\times 46,294) = -6,200,348$ and the BIC is $-5,604,714$.

All the offset parameters applying to visible peaks are non-zero so none of the peaks in the range are symmetrical. This was also true for the majority of the peaks in the other parts of the dataset. This suggests that the assumption of peaks having longer right hand tails was sensible. The value of the offset peak $\mu_8 + \delta_8$ given in table 3.13 has moved 100 Da to the right compared with μ_8. This is the upper limit of the permissible values for a δ. We can see from figure 3.16 that the relative intensity at both the m/z values is negligible and so this causes no problem. In the final model the variance scaling parameters ω seem to be around the value 2 which suggests our previous model using a standard scaling of 2 was not unreasonable.

No patterns were visible in any of the trace plots for the parameters so the chains were acceptable. To check if any peaks were different between the two stages of melanoma t-statistics were again calculated using the MAP estimates of θ at each peak location and then correcting for multiple testing. The locations remain essentially the same as in the previous two models and the majority of the t-statistics are positive showing that the heights of the peaks in stage IV are lower than in stage I. However, at around 11,500 to 11,900 Da the t-statistics show the opposite. In Mian et al. (2005) the area around 11,701 Da was identified as one showing significant variability in the data. The absolute value of the majority of t-statistics increases under the two offset peaks models.

3.6 Summary

In this chapter we have shown how MCMC algorithms can successfully be used to simulate from a model for mass spectrometry data and also how we can incorporate the peak finding procedure from chapter 2 to provide a suitable starting point. The use of these methods greatly reduces the dimension of the datasets to a relatively small number of parameters. The number of datapoints are 2,009,088 and 2,859,955 for the breast cancer and melanoma datasets respectively. For the final model in this chapter the respective numbers of modelled parameters were 40,192 and 46,294 - around 2% of the original number of datapoints in each case.

It has been shown that it is important to consider the data not as a combination of single peaks but as a combination of double peaks with offset locations. Using the AIC calculations it is seen that using offsets to model the peaks gives much better results. The MAP curves match the data more closely for only a slight increase in the number of parameters. A summary of the AIC results is shown in table 3.15. We conclude that although the AIC is lowest for the different variances model, it is not a large amount lower than the AIC for the simpler offset peaks model for either dataset when compared with the reductions for the previous model alterations. To model the data we should use offset peaks with different variances.

	cancer			melanoma		
	AIC	BIC	# parameters	AIC	BIC	# parameters
single	-2,594,632	-2,343,817	20,044	-5,576,214	-5,278,873	23,110
double	-3,090,994	-2,591,523	39,916	-5,725,350	-5,132,598	46,070
offset	-4,283,938	-3,782,705	40,056	-6,181,296	-5,587,103	46,182
variance	-4,296,174	-3,793,243	40,192	-6,200,348	-5,604,714	46,294

Table 3.15: The AIC and BIC statistics for the four models considered in this chapter.

For comparison we will now also consider the BIC statistics. From the BIC results we see that, in agreement with the previous conclusions using AIC, the model including offset peaks with differing variances is deemed to be the most suitable. When moving from the single peaks to the double peaks model, the BIC heavily penalised the introduction of nearly double the number of parameters for the melanoma dataset which led to an increase in the statistic. However, when the more complex models were analysed the BIC statistic fell again.

In this chapter we split the datasets into sections and model each of the sections separately. This resulted in having around 20 different estimates for the proportionality constant ξ - one in each section instead of the overall parameter that we would prefer to model. When checking the values of ξ obtained for each section it was found that, for all sections that contained visible peaks, the value of ξ converged to roughly the same value in each section. For the sections without visible peaks the value of ξ was larger but the peaks had negligible height. It seems reasonable to assume that the value of ξ is approximately constant over all sections.

The analysis of the breast cancer and melanoma datasets using the MCMC methods discussed in this chapter requires a large amount of computational time. This is primarily because each section of the data must be run sequentially so as not to split processor time between tasks. The High Performance Computing (GRID) system was used to compare the times taken to analyse the datasets. Using a desktop computer to carry out the analysis of the complete datasets resulted in total analysis times of 563.5 minutes and 706.5 minutes for the breast cancer and melanoma datasets respectively. When using the GRID, the time taken for the analysis of each section of data was reduced by approximately 25% in both datasets. However, since there are multiple processors on the GRID, each section of the data can be submitted to a different one and the analysis carried out in parallel. This reduces the time taken for each dataset to the time taken for the largest section. The times taken when running the analysis in parallel are 74.9 minutes and 78.4 minutes for the breast cancer and melanoma datasets respectively.

In Dryden et al. (2005) the breast cancer dataset was analysed using a variant of the Hotelling T^2 test (Hotelling, 1931). The day information can be taken into account when using the Hotelling test as vectors can be constructed of the peak heights over all four days. A brief description of the technique now follows.

Let $\bar{\mathbf{x}}_{Ai}$ and $\bar{\mathbf{x}}_{Bi}$ be the q-vectors of means in groups A and B respectively at m/z value i $(i = 1 \ldots p)$, with sample sizes n_A, n_B. Let $\mathbf{S}_{xi}$ be the unbiased pooled within-group $q \times q$ covariance matrix at m/z value i. For the breast cancer data $q = 4$ as there are 4 days of information available and $p = 13951$ is the number of recorded m/z values between 2000 and 30,000 Da. The two sample Hotelling T^2 test of $H_0 : \mu_{Ai} = \mu_{Bi}$ versus $H_1 : \mu_{Ai} \neq \mu_{Bi}$ at m/z value i is $T^2_{x,i} = (\bar{\mathbf{x}}_{Ai} - \bar{\mathbf{x}}_{Bi})^T \mathbf{S}^{-1}_{xi} (\bar{\mathbf{x}}_{Ai} - \bar{\mathbf{x}}_{Bi})$ under certain assumptions (see Dryden et al., 2005) and we reject H_0 in favour of H_1 at the $100\alpha\%$ level if

$$T^2_{x,i} > T_{crit}(\alpha) = \frac{(n_A + n_B)(n_A + n_B - 2)q}{n_A n_B (n_A + n_B - q - 1)} F_{q, n_A + n_B - q - 1}(1 - \alpha)$$

where $F_{\nu_1,\nu_2}(1-\alpha)$ is the $1-\alpha$ quantile of the F_{ν_1,ν_2} distribution.

The Dryden et al. (2005) method tries to account for the extra noise which would be inherent in further repetitions of the experiment. The noise is considered to be *iid* Gaussian with mean zero and variance σ^2 and thus the unobserved noisy vector $w_i = x_i + \epsilon_i$ where $\epsilon_i \sim N_q(0, \sigma^2 \mathbf{I}_q)$ independently. The offset test statistic is then $T^2_i(\sigma^2) = (\bar{\mathbf{x}}_{Ai} - \bar{\mathbf{x}}_{Bi})^T (\mathbf{S}_{xi} + \sigma^2 \mathbf{I}_q)^{-1} (\bar{\mathbf{x}}_{Ai} - \bar{\mathbf{x}}_{Bi})$. Given σ^2, $T^2_i(\sigma^2)$ can be observed and these statistics can be used for inference. A suitable value of σ^2 is determined by a calibration method and subsequently H_0 is rejected if $T^2_i(\sigma^2) > T_{crit}(\alpha)$.

The significant results obtained from this analysis are shown in table 3.16. When comparing these results with the MCMC result tables presented in this chapter we see that most of the values in table 3.16 are identified by the MCMC method. The exceptions are the rows including 7,687 Da, 11,381 Da and 15,377 Da for the control/treated comparisons and the rows including 6,231 Da, 6,552 Da, 10,169 Da and 13,811 Da for the control/control comparisons.

adc/adt	tdc/tdt	mcc/mct	adc/tdc	adc/mcc	tdc/mcc
					3839
				4127	4120
				4396	4396
					4648
				4813	4798
					5364
		5692	5653	5661	
			6231	6282	
			6552	6552	
7029			7017	7019	
7687			7685		
				8094	
			10169		
		10265			10248
		11381	11351	11369	11340
		13854	13811		13831
		14028	14048	14055	
15377			15390	15402	

Table 3.16: Significant *m/z* values in the breast cancer dataset from Dryden et al. (2005). Similar values are listed on the same line.

There are some differences between the results from the two methods. In the MCMC analysis the significant result at 7,029 Da is between *mccon* and *mctax* - in table 3.16 it is between the *adcon* and *adtax* groups. Also the last row in table 3.16 shows differences with adc/tdc and adc/mcc - in the MCMC analysis the difference at this *m/z* value is with tdc/mcc comparison.

It should be noted that the Hotelling analysis does not reveal any significant differences between any of the spectra for *m/z* values higher than 15,500

Da. The MCMC approach provides many such *m/z* values of which the majority are for mcc/mct and tdc/mcc comparisons.

In Mian et al. (2003) the breast cancer dataset was studied using *artificial neural networks* (ANNs). This research highlighted the *m/z* values 10,518 Da, 11,100 Da, 11,687 Da and 13,239 Da as showing good classification ability between control and treated cell-lines. Only *m/z* values between 10 kDa and 15 kDa were considered in that analysis. The models described in this chapter do not reproduce any of these results. However, since the work was investigating classification and we are testing for significance one may expect different conclusions.

In Mian et al. (2005) the melanoma dataset was studied using ANNs. They find that there was a large difference between stage I and stage IV melanoma at 11,700 Da in terms of the variability between the two groups at this *m/z* value. The models used in this chapter do identify this *m/z* value as one exhibiting differences between the two groups.

Chapter 4

Applying Multilevel Modelling to Proteomic Mass Spectrometry Data

4.1 Introduction

In section 1.5 an overview of some multilevel modelling techniques was presented. This chapter will consider the use of this framework to model the available datasets via the use of mixed effect models. Inference will be carried out using the software package MLwiN (Rasbash et al., 2000). The construction of the model will be described in section 4.2 and the results obtained from the analysis are described in section 4.3. A series of related models are considered and the results are compared with those from the MCMC analysis presented in chapter 3.

4.2 The Model

One of the main aims of this thesis is to identify locations along a spectrum which enable us to differentiate between drug-treated breast cancer cell-lines and non-treated controls and to differentiate between stage I and stage IV melanoma. In this chapter we aim to find these locations by considering a multilevel model for the data. Multilevel modelling is a useful technique to apply when the data have an obvious hierarchical structure such as here where we have m/z values within spectra.

4.2.1 Model Summary

In chapter 3 the data were considered as a mixture model of Gaussian peaks. The initial model used was $y_{is} \sim N(\theta_{is}, \tau^{-1})$ where θ_{is} is

$$\theta_{is} = \sum_{j=1}^{k} h_{js}(\xi\mu_j^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\xi\mu_j^2)^{-1}(x_i-\mu_j)^2\right). \tag{4.1}$$

The index $i = 1, \ldots, p$ represents the position on the spectrum, x_i is the i^{th} m/z value, $s = 1, \ldots, n$ is the spectrum number, $j = 1, \ldots, k$ is the peak number, ξ is the constant of proportionality and μ_j are the peak locations.

A more complex model was then introduced which considered each peak in the data as a combination of two peaks - one spiky peak as already modelled and another with a larger variance to accommodate a baseline amount of intensity. This larger variance was set to be twice the variance of the spiky peak. Offset peaks were also considered in a third model which resulted in better matching to the data for asymmetrical peaks. The equation for θ_{is} for the double and offset peaks models is

$$\begin{aligned}\theta_{is} = & \sum_{j=1}^{k} h_{js}(\xi\mu_j^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\xi\mu_j^2)^{-1}(x_i-\mu_j)^2\right) \\ & + \sum_{j=1}^{k} h_{js}^{*}(\omega_j\xi(\mu_j+\delta_j)^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\omega_j\xi(\mu_j+\delta_j)^2)^{-1}(x_i-(\mu_j+\delta_j))^2\right)\end{aligned} \tag{4.2}$$

where i, x_i, s, n, j, ξ and μ_j are as in equation (4.1), the δ_j parameters model the offset of the second peaks from the main location μ_j (set to zero for the double peaks model) and the ω_j are the variance scaling factors for the double peaks. For the double peaks and offset peaks models these ω_j were set to 2. A final extension of the model considered the possibility that the variance scaling parameters ω_j were different for each peak.

4.2.2 The Multilevel Model

In the two datasets to be analysed there exists a two-level structure - *m/z* values within spectra. As shown in section 1.6.3 a multivariate normal model is of the form $\mathbf{y} \sim MVN(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ and for two-level models this can be rewritten as

$$y_{ij} = \mathbf{X}_{ij}^T\boldsymbol{\beta} + \mathbf{Z}_{ij}^T\mathbf{u}_j + e_{ij}$$

$$\mathbf{u}_j \sim MVN(0, \boldsymbol{\Omega}_u)\ ,\ e_{ij} \sim N(0, \sigma_e^2)$$

where i indexes *m/z* value and j indexes spectrum in our example. The value of y is the observed datapoint at *m/z* value i in spectrum j, $\boldsymbol{\beta}$ and $\mathbf{u}$ are vectors of parameters, the matrix $\mathbf{X}$ is the design matrix for the fixed effects and the matrix $\mathbf{Z}$ is the design matrix for the random effects. The $\mathbf{u}_j$ and the e_{ij} are independent. Such a model is called a linear mixed effect model.

As discussed previously the main points of interest in the data are the location and height of peaks and from this information we wish to discover where the groups are different. We can accommodate this requirement in the model by fitting a common fixed effect for each peak in each group to represent an average intensity. Differences in the fixed effect estimates will show any locations with significant differences in peak height between groups. Incorporating random effects for each peak in each spectrum will better match the fitted model to the data.

Due to computational limitations, inference on the full dataset is difficult in MLwiN. In order to analyse the data we must therefore consider using the splitting algorithm described in section 3.2.6 which creates a partition of the available *m/z* values into a number of sections. We obtain 17 sections for the breast cancer dataset and 19 sections for the melanoma dataset.

We now describe how to create the fixed and random predictors used in the initial model of equation (4.1). This model does not involve the presence of multiple peaks at each location. Firstly the peak finding method described in chapter 2 is used to obtain 150 peak locations across the whole dataset. For each section of data the relevant peak locations are selected in order from this list so that they are in order of decreasing peak size. Consider a section of the data of length t with p relevant peak locations. For the first peak location, μ_1, identified in this section by the algorithm we construct a peak at this location from the equation

$$\theta_i \;=\; (\xi\mu_1^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\xi\mu_1^2)^{-1}(x_i-\mu_1)^2\right) \tag{4.3}$$

where the index $i = 1, \ldots, t$ represents the position on the spectrum, x_i is the i^{th} m/z value and ξ is the constant of proportionality. The value of ξ is obtained from the MCMC analysis in chapter 3.

We now create from this the design matrices for the **X** and **Z** predictor variables associated with the first peak location. The random predictor consists of n replicates of this first fitted peak where n is the number of spectra in the dataset. If there are g groups in the dataset (6 for the breast cancer and 2 for the melanoma) then we also create g fixed predictors for this first peak. The parts of the fixed and random design matrices relevant to the first group are identical. The fixed predictors for the remainder of the groups indicate where we wish to measure differences from the first group. Fixed and random predictors should be created this way for each of the peak locations identified by the algorithm. For an example structure diagram for the fixed effect matrix see figure 4.1. In this figure we have a dataset containing three spectra - one in each of three groups - and two peak locations, the second earlier (in m/z terms) than the first.

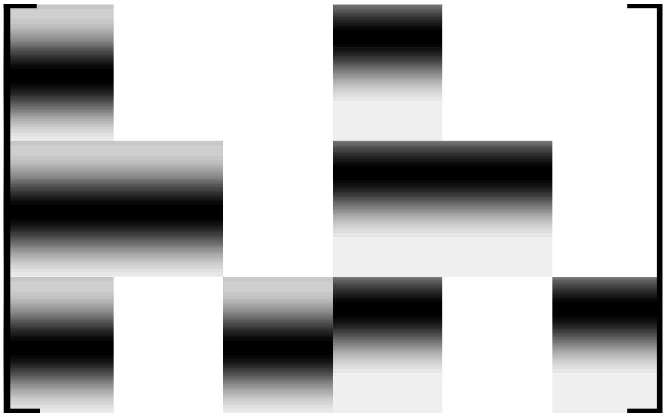

Figure 4.1: Diagram showing the general structure of the fixed effects design matrix **X**. The matrix is of dimension 3t × 3p. The areas of white contain zeros and the gradient of grey represents the slopes of the peak. The top of the peak is indicated by black.

For the breast cancer dataset we have 144 spectra in 6 groups so the dimension of the fixed effect design matrix **X** is $144t \times 6p$ where t is the length of the section of data and p is the number of peaks fitted. The respective dimension for the melanoma dataset is $205t \times 2p$.

For the more complex models we calculate additional fixed and random effects using the same method. These extra effects are to model the presence of the double peaks at each location. For the j^{th} peak location we construct the peak using the equation

$$\theta_{is} \;=\; (\omega_j\xi(\mu_j+\delta_j)^2)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\omega_j\xi(\mu_j+\delta_j)^2)^{-1}(x_i-(\mu_j+\delta_j))^2\right) \tag{4.4}$$

where, in addition to the parameters in equation (4.3), δ_j models the offset of the double peak from the original location (zero for the double peaks model) and ω_j models the scaling parameter for the double peak variance (2 for the double and offset peaks models).

In summary, if the total number of peaks fitted in the model is p and the number of spectra in the dataset is n then the model becomes

$$\underset{nt\times 1}{\mathbf{Y}} = \underset{nt\times gp}{\mathbf{X}}\,\underset{gp\times 1}{\boldsymbol{\beta}} + \underset{nt\times pn}{\mathbf{Z}}\,\underset{pn\times 1}{\mathbf{U}} + \underset{nt\times 1}{\mathbf{E}}$$

$$\mathbf{U}\sim MVN(\mathbf{0},\boldsymbol{\Omega}_u)\ ,\ \mathbf{E}\sim N(\mathbf{0},\boldsymbol{\Omega}_e).$$

The vector $\mathbf{Y}$ is the vector obtained by stacking the spectra one by one into a single column and the vector $\mathbf{E}$ is the stacked error vector. The first column of the fixed effect design matrix $\mathbf{X}$ consists of n replicates of the first fitted peak. The second column to the g^{th} column, where g is the number of groups, consist of replicates of the first fitted peak when the corresponding points in the $\mathbf{Y}$ vector are in groups $2,\dots,g$ and zero otherwise. This pattern repeats in the next columns of $\mathbf{X}$ for other peaks in the model. The random effects design matrix $\mathbf{Z}$ is similar to the fixed effects design matrix $\mathbf{X}$ except we do not distinguish between groups. However, we do distinguish between spectra and a different random effect will be obtained for each spectrum. Therefore the i^{th} column of the $\mathbf{Z}$ matrix consists of n replicates of the i^{th} fitted peak.

The model parameters $\boldsymbol{\beta}$ and $\mathbf{U}$ are estimated by using an iterative procedure, namely the iterative generalised least squares (IGLS) algorithm as described in section 1.6.4.

4.2.3 The Shifting Procedure

The peak finding procedure from chapter 2 finds a series of locations where Gaussian peaks can be placed which allow us to obtain a quick approximation of the data. However, we found in chapter 3 that the locations obtained from this algorithm may not be 'optimal' in the sense that the model deviance could be decreased by adjusting the peak locations.

To reduce the deviance for the mixed effect models investigated in this chapter we can consider shifting the location of a peak by one recorded m/z value at a time in either direction until we reach a local minimum deviance. This involves the recalculation of the design matrices $\mathbf{X}$ and $\mathbf{Z}$. If the deviance increases after the first shift we should instead consider moves in the opposite direction until we reach the local minimum. This procedure should be used on each of the peaks in the model. Every time a peak is moved the fixed and random predictors associated with that peak will need to be recalculated.

For the case where two or more peaks are close together we should repeat the procedure on these peaks. It is possible that, after moving the first peak, the movement of the second one has an effect on the best location for the first. After all the peaks have been moved once we should check that we cannot decrease the deviance further by making extra changes. If peaks are far apart then this second round of checking should not be required.

4.3 Application to Datasets

4.3.1 Results for the Breast Cancer Dataset

Firstly the IGLS algorithm was run to get estimates for each of the fixed and random effects and the variances Ω_u and σ_e using the 150 peaks obtained from the peak finding algorithm. To try and improve the fit of the model the shifting procedure described in section 4.2.3 was used. The first peak was moved one m/z value at a time to the left or right until the deviance did not decrease further. This procedure was repeated for all of the peaks in the model. After carrying out this shifting it appeared that the peaks had moved to locations similar to those obtained from the MCMC analysis in chapter 3.

In figure 4.2 the original data from 7,600 Da to 8,200 Da are shown along with the fixed effects which model the average spectrum for the groups. The black line shows the fixed predictor under the single peaks model and the blue line shows the fixed predictor under the model where the double peaks can have offset peaks and differing variances. The change in fit is most obvious for the peak around 8,100 Da. Here the peak is non-Gaussian and the increase in model complexity has allowed the model to better fit the peak shape. In the groups *adcon*, *tdcon* and *tdtax* there there are two obvious peaks present between 7,800 Da and 8,000 Da. The fixed effect in these groups changes to model the data more accurately. The peak at 7,700 appears to be fairly symmetric so the fixed effect does not change greatly between the two models shown.

To determine whether any of the peaks could be removed from the model Z-tests were calculated at the 5% level. If the fixed effects for a peak were insignificant for all of the six groups then the peak was removed. This resulted in 22 peak locations being omitted which leaves 128 to model the presence of a peak in any of the six groups.

These 128 remaining locations at which a peak was present were further checked to identify where the fixed effects significantly differed between groups. As in chapter 3 we need to correct for the large number of tests being carried out. This was achieved by using the false discovery rate algorithm described in section 1.6.6 with $q^* = 0.05$. Following this procedure, forty five peaks were deemed significant for one or more of the control/control comparisons and their

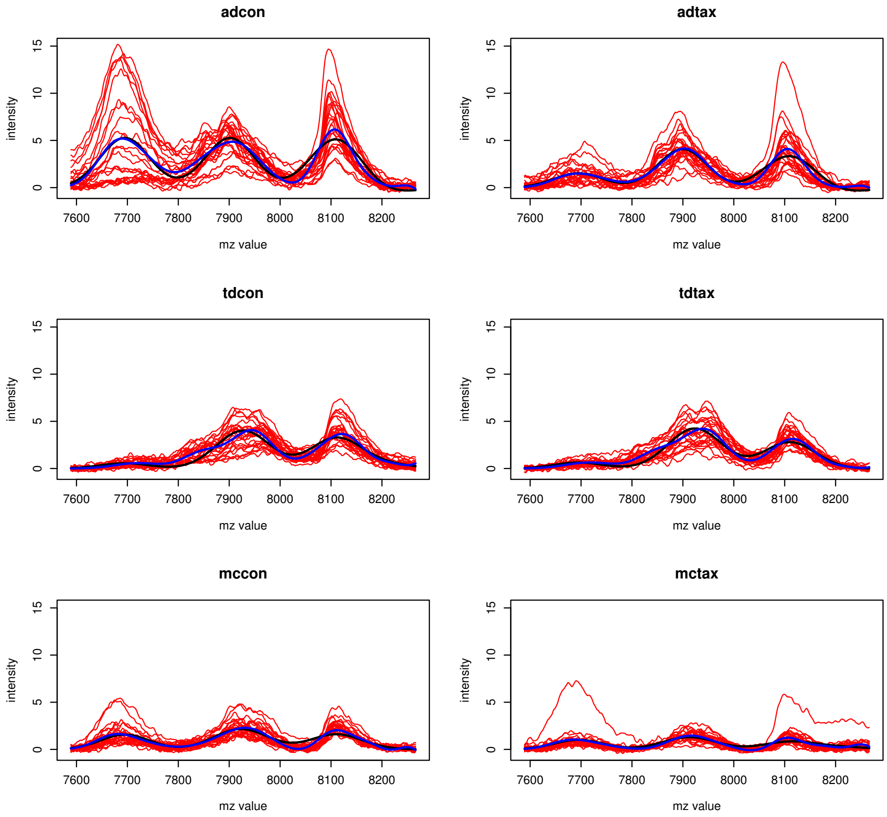

Figure 4.2: Plots of the breast cancer data with the fixed effects under the single peaks model (black) and the differing variance model (blue).

locations are shown in table 4.1. Thirty peaks were significant for one or more of the control/treated comparisons and their locations are shown in table 4.2. From table 4.2 it can be seen that there are some differences identified between the two chemoresistant groups adcon and adtax. The treatment should have no effect on these cell lines. Also only one location was identified in the MCMC analysis in chapter 3 which exhibited differences between these two groups. However, in chapter 3 the tests for significance were separated by day and this information was not considered in the multilevel framework considered in this chapter. This may explain some of the differing results between the approaches.

m/z	2226	2426	2724	2954	3584
groups	AM	AT,TM	AT,AM	TM	AM,TM

m/z	3709	4019	4229	4329	4389
groups	TM	AT,TM	AT	AT	AM,TM

m/z	4641	4703	4803	4887	5103
groups	AM,TM	AM,TM	AM,TM	TM	AT,TM

m/z	5376	5553	5669	6566	7019
groups	AT,TM	AT	AT,AM	AT	AT

m/z	7146	7274	7694	7939	8110
groups	AT,TM	AT	AT,AM	AT	AM

m/z	8212	9187	9651	10115	10230
groups	AT	TM	TM	TM	AT,TM

m/z	10435	10888	11137	11357	11918
groups	AM,TM	AT	AT,AM	AT	AT,AM

m/z	12643	13432	14046	14250	14857
groups	AM	AT	AT	AT	AT,TM

m/z	15402	16315	17764	20360	26743
groups	TM	AT	AT,TM	AT	AT

Table 4.1: The peak locations in the breast cancer dataset with significant differences between adcon/tdcon(AT), adcon/mccon(AM) and tdcon/mccon(TM).

m/z	2186	2687	3590	4391	4703
groups	T,M	T,M	M	T	T

m/z	5406	6693	7019	7694	8110
groups	A,M	T	M	A	A,M

m/z	8473	8769	10228	10426	11718
groups	T	M	M	M	A,T

m/z	12643	13432	14046	14857	15043
groups	A,M	A	M	M	T

m/z	15402	16315	17983	19030	20010
groups	M	A	T	M	A,T

m/z	20815	21010	23083	25373	26743
groups	M	A	T	M	A

Table 4.2: The peak locations in the breast cancer dataset with significant differences between adcon/adtax(A), tdcon/tdtax(T) and mccon/mctax(M).

The fixed effect parameters for the adcon group appeared to be larger than for the other groups at most peak locations. One notable exception occurs at an *m/z* value of 4,389 Da. This location had two of the most significant differences between groups of all the peaks modelled. These occurred between the *adcon* and *mccon* groups and between the *tdcon* and *mccon* groups. Plots of the original data around this *m/z* value are shown in figure 4.3. Also highly significant were the peaks at *m/z* values of 10,231 Da and 10,425 Da. At both of these locations the difference lies between the *tdcon* and *mccon* groups.

Plots of these m/z values are shown in figure 4.4.

The locations 4,389 Da and 10,231 Da identified here as exhibiting some of the most significant differences between groups were also identified as the same in the MCMC analysis presented in chapter 3. In addition they were identified in the classification analysis presented in chapter 2 as being some of the best classifiers of new spectra. The remainder of the top ten most significant locations identified by the multilevel analysis also appear in the MCMC analysis results shown in tables 3.4 and 3.5 but not in the classification results. This lack of similarity should not be unexpected as we are examining classification in chapter 2 and identifying differences in chapter 4.

For the MCMC analysis we concluded that it was important to consider the peaks as a combination of two offset peaks with different variances as the AIC statistic was lowest for this model. We can incorporate this into the multilevel model by changing the fixed and random effects appropriately. For each of the three more complex models (double peaks, offset peaks and differing variance peaks) the number of fixed and random effects will double compared with the original model to accommodate the parameters for the peaks with larger variances.

In figure 4.5 the estimates of the data are shown for each of the four models considered in this chapter for the m/z values between 7,400 Da and 8,400 Da. The data are reflected in the x-axis for comparison. The single peaks model is represented by the red curve, the double peaks model by the black curve, the offset peaks model by the green curve and lastly the differing variances model by the blue curve.

At 8,100 Da and 7,950 Da the fit to the data improves as the complexity of the model increases, with the blue curve providing a visibly better match. At around 8,050 Da moving through the four models (red to blue) allows the

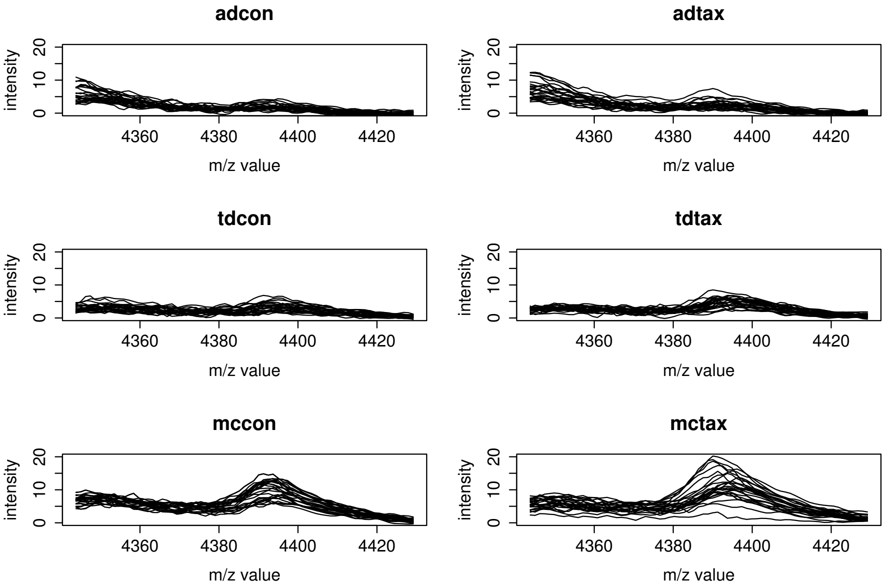

Figure 4.3: Plots of the original data around the *m/z* value 4,389 Da separated by group.

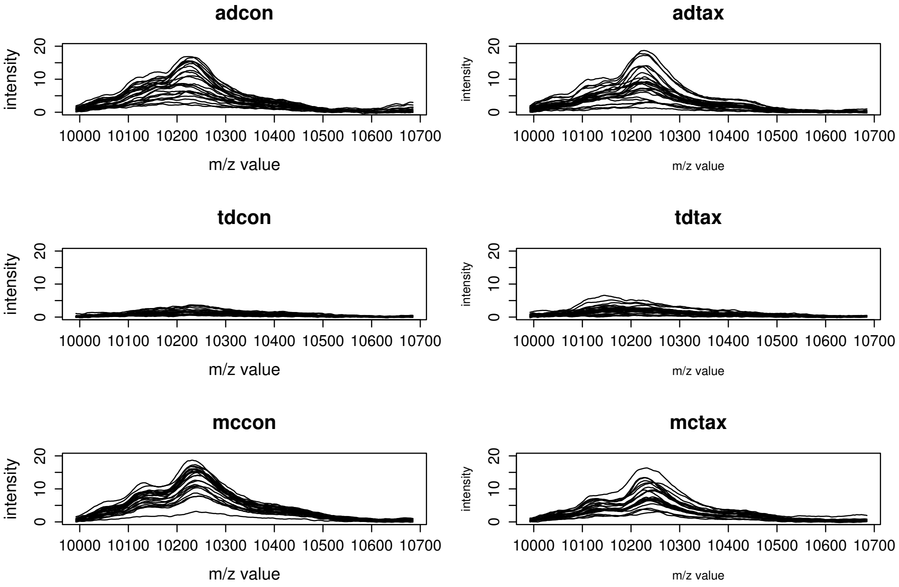

Figure 4.4: Plots of the original data around the *m/z* values 10,231 Da and 10,425 Da separated by group.

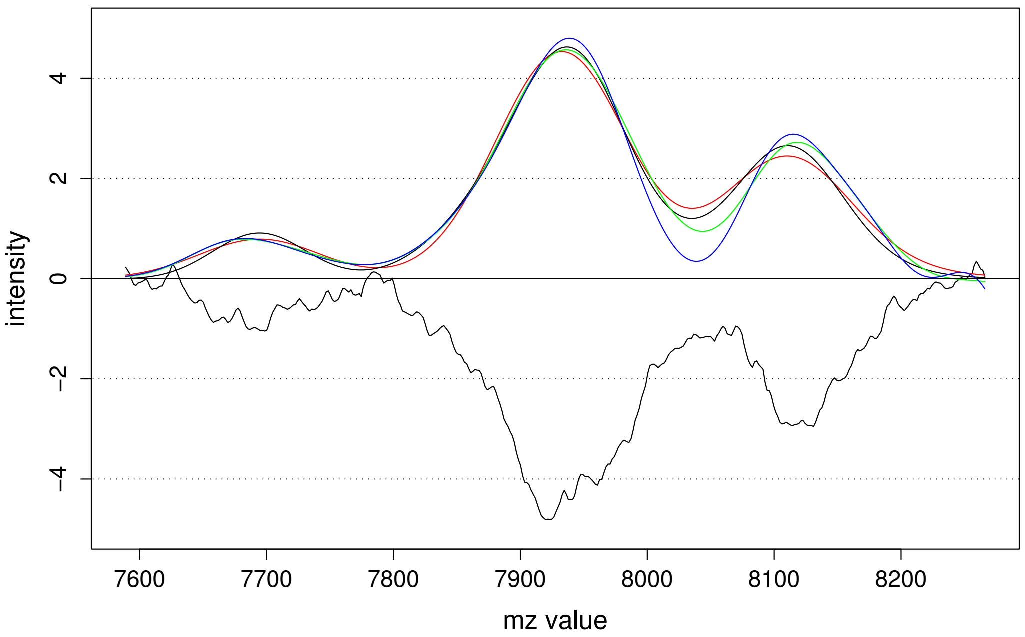

Figure 4.5: The estimates obtained for one spectrum from the single(red), double(black), offset(green) and differing variance(blue) models as compared to the original breast cancer data (reflected in the x axis).

lowest point between the two peaks to be reduced. Modelling using offsets (green and blue) has allowed the non-symmetric peak at 8,050 Da to be more closely modelled and also the peak at 7,700 Da to move to the left resulting in a better fit to the data. The deviances associated with the four models are shown in table 4.3.

The single peaks and double peaks models are nested and so a likelihood ratio test can be calculated to determine which of the two models is preferable. The number of parameters is increased by 896 when moving to the double peaks model and these parameters consist of 768 extra fixed effects and 128 extra

	model			
	single	double	offset	variance
deviance	-1,259,241	-1,563,974	-1,796,832	-1,921,647

Table 4.3: The deviances for the four models of the breast cancer data considered in this chapter.

variances. The test is thus

$$-1,259,241 - (-1,563,974) = 304,733 >> \chi^2_{896,0.95} = 966.8$$

and the double peaks model is a significant improvement on the simpler model. The offset and differing variances models are not nested and so likelihood ratio tests cannot be used. However, the AIC and BIC statistics can be calculated and are shown in table 4.4. The number of parameters does not change after the double peaks model and so we find that the the model considering peaks with differing variances is the most preferable as it has the lowest deviance.

	model			
	single	double	offset	variance
AIC	-1,257,449	-1,560,388	-1,793,246	-1,918,061
BIC	-1,246,237	-1,537,951	-1,770,809	-1,895,624

Table 4.4: The AIC and BIC statistics for the four models of the breast cancer data considered in this chapter.

4.3.2 Results for the Melanoma Dataset

Before starting the multilevel analysis the peak finding algorithm from chapter 2 was implemented to obtain 150 peak locations across the entire dataset. These locations were used to create the fixed and random predictor matrices and the IGLS algorithm was then used to obtain estimates for the effects and the variances Ω_u and σ_e for each section of the data. Using the shifting procedure described in section 4.2.3, each peak was moved one m/z value at a time to the left or right in order to minimise the deviance of the model. This procedure was repeated for all of the peaks in the model and, after the procedure was completed, the peaks had moved to locations similar to those obtained from the MCMC analysis in chapter 3.

In figure 4.6 the original data from 7,400 Da to 8,500 Da are shown along with the fixed effects which model the average spectrum for the groups. The black line shows the fixed effect under the single peaks model and the blue line shows the fixed effect under the model where the double peaks can have differing variances. The change in fixed effect is most obvious in the stage I plot. As the peak at 7,800 Da is non-symmetric, the addition of peak offsetting has enabled the fitted peak to change shape to better match the original data. The peak height in stage IV remains roughly the same although the location changes, whereas both the height and location change in stage I.

To determine whether any of the peaks could be removed from the model Z-tests were calculated at the 5% level. If the fixed effects for a peak were insignificant for both stage I and stage IV then the peak was removed. This resulted in 46 peak locations being removed which leaves 104 to model the presence of a peak in one or both of the groups. These remaining locations were studied to identify where the fixed effect in stage I was significantly different to that in stage IV. After correcting for multiple testing using the false discovery rate algorithm in section 1.6.6, forty peaks were deemed significant and their locations are shown in table 4.5 along with the associated values of the Z

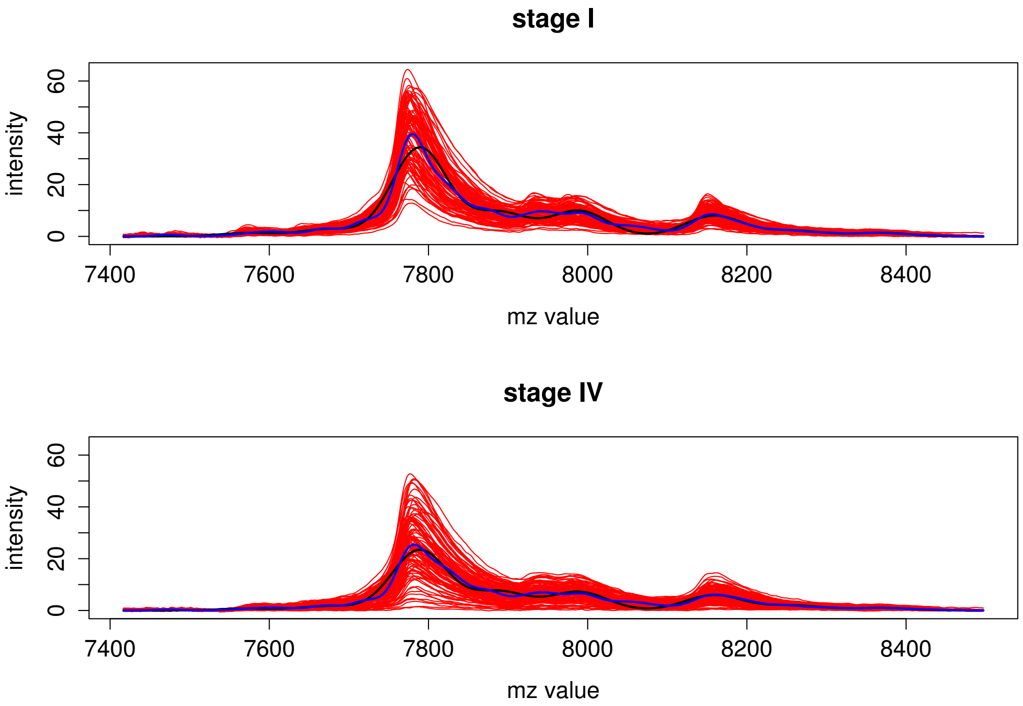

Figure 4.6: Plots of the melanoma data with the fixed effects under the single peaks model (black) and the differing variance model (blue).

statistic. From table 4.5 it can be seen that the fixed effect parameters for stage IV generally appeared to have lower values than their counterparts in stage I. There are only 3 m/z values where the peaks have larger magnitude in stage IV - 6,754 Da, 11,524 Da and 11,705 Da. For the two peaks in the 11,600 Da area the fixed effect in stage I is deemed not significantly different to zero and so these locations refer to peaks only present in stage IV. For the peak at 6,754 Da the fixed effects in both groups were significantly different to zero and thus there is a peak present in both groups. The original data at the two m/z values around 11,600 Da are shown in figure 4.7 showing the obvious difference between stage I and stage IV.

m/z	2227	2312	2489	2535	2729
Z statistic	-8.250	-6.000	-8.333	-5.167	-6.000

m/z	2777	2960	3065	3304	3552
Z statistic	-6.333	-8.333	-8.250	-5.889	-6.667

m/z	3891	3972	4477	4653	4777
Z statistic	-8.833	-4.545	-4.333	-5.000	-5.250

m/z	6452	6650	6754	7571	7671
Z statistic	-5.890	-5.532	6.469	-8.000	-7.538

m/z	7789	7986	8156	8364	8931
Z statistic	-7.438	-5.690	-5.606	-5.000	-4.750

m/z	9323	9481	9672	11524	11705
Z statistic	-4.941	-5.452	-4.500	5.636	5.469

m/z	13943	14210	14655	17235	17424
Z statistic	-5.147	-7.230	-6.000	-4.688	-5.200

m/z	17811	18206	28170	28742	29239
Z statistic	-5.143	-4.750	-7.543	-7.125	-6.600

Table 4.5: The peak locations with significant Z statistics between the two stages of melanoma.

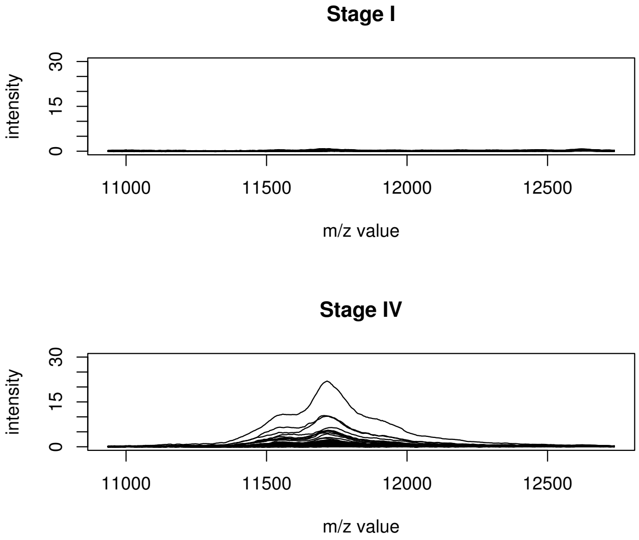

Figure 4.7: Plots of the melanoma data at the peak locations where the stage IV intensity is higher than in stage I.

In chapter 3, considering the peaks as a combination of two offset peaks with different variances was deemed important because this model had the lowest AIC statistic. By changing the fixed and random predictors appropriately this structure can be incorporated into the multilevel model. For each of the three more complex models (double peaks, offset peaks and differing variance peaks) the number of fixed and random effects increases from 208 and 104 respectively under the single peaks model to 416 and 208 to accommodate the parameters for the peaks with larger variances.

In figure 4.8 the estimates of the data are shown for each of the four models considered in this chapter for the m/z values between 7,400 Da and 8,400 Da. The data are reflected in the x-axis for comparison. The single peaks, double, offset and differing variances models are respectively represented by the red, black, green and blue curves.

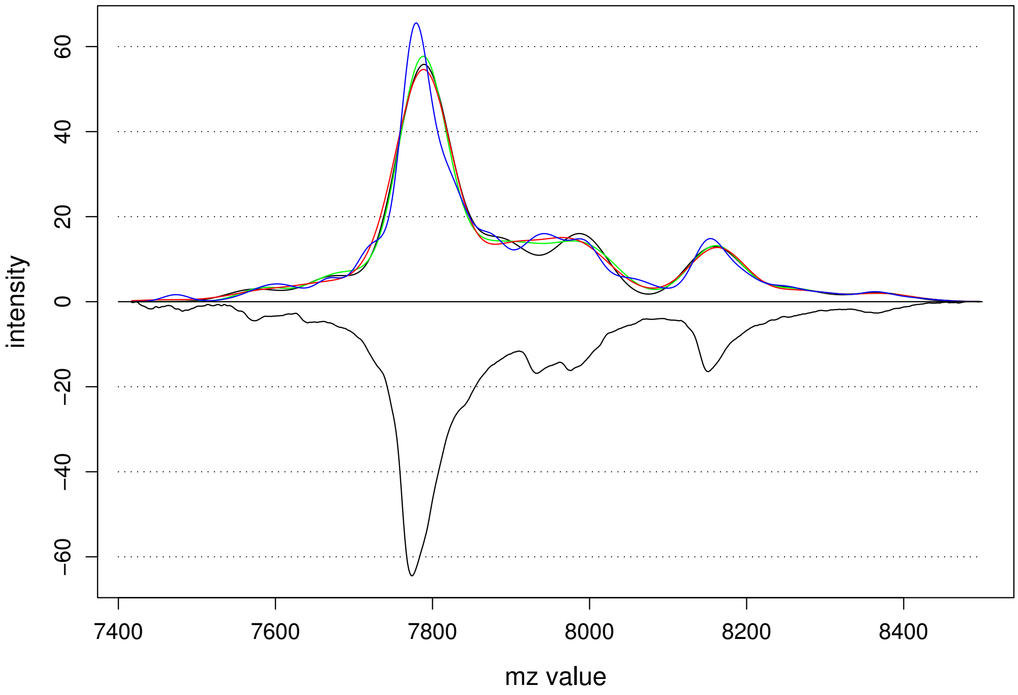

Figure 4.8: The estimates obtained for one spectrum from the single(red), double(black), offset(green) and differing variance(blue) models as compared to the original melanoma data (reflected in the x axis).

At 7,800 Da and 8,150 Da increasing the complexity of the model provides a much better fit to the data, with the blue curve showing the best match of the four models. At around 7,950 Da moving from the single to the double peaks model (red to black) allows the peaks to be modelled more closely because of the reduction in the ξ parameter. In that area there are now troughs present instead of just an overall curve as there was under the single peaks model. Although the second peak around 7,950 Da is modelled more closely with the black curve the first peak in that area is not fitted well. When the additional complexity of the offset peaks and different variances is implemented the fit is much improved. The deviances associated with the four models are shown in table 4.6.

	model			
	single	double	offset	variance
deviance	-2,228,228	-2,527,056	-2,801,919	-2,980,246

Table 4.6: The deviances for the four models of the melanoma data considered in this chapter.

Since the single peaks and double peaks models are nested, we can carry out a likelihood ratio test to determine which of the two models is preferable. The number of parameters is increased by 312 when moving to the double peaks model (208 extra fixed effects and 104 extra variances) and thus the test is:

$$-2,228,228 - (-2,527,056) = 298,828 >> \chi^2_{312,0.95} = 354.2$$

and the double peaks model is thus a significant improvement on the simpler model. Likelihood ratio tests cannot be used for further model comparisons as the offset and differing variances models are not nested. Using the model deviances and the AIC/BIC statistics shown in tables 4.6 and 4.7 we consider the differing variances model to be the most preferable as it has both the lowest deviance and the lowest AIC/BIC.

	model			
	single	double	offset	variance
AIC	-2,227,602	-2,525,806	-2,800,669	-2,978,996
BIC	-2,223,574	-2,517,764	-2,792,627	-2,970,954

Table 4.7: The AIC and BIC statistics for the four models of the melanoma data considered in this chapter.

4.4 Summary

In this chapter we have shown how multilevel modelling techniques can successfully be used to model mass spectrometry data. The peak finding procedure from chapter 2 is used to provide a suitable starting point for the fixed and random predictors. These predictors are then shifted by one recorded *m/z* value at a time in order to minimise the deviance of the model. Using mixed effect models means that the data are described using a much smaller number of parameters than there are original datapoints.

In agreement with the results from chapter 3 it has been shown that it is important to consider the data not as a combination of single peaks but as a combination of double peaks with offset locations. Using likelihood ratio tests where applicable and AIC/BIC calculations otherwise it has been shown that using the model allowing offset peaks with different variances gives much better results than the simpler models. By doubling the number of parameters in the model we obtain a much better match to the data. A summary of the model deviances and AIC/BIC statistics are shown in table 4.8. We conclude that the most complex of the four models is the most preferable.

	cancer			melanoma		
	deviance	AIC	BIC	deviance	AIC	BIC
single	-1,259,241	-1,257,449	-1,246,237	-2,228,228	-2,227,602	-2,223,574
double	-1,563,974	-1,560,388	-1,537,951	-2,527,056	-2,525,806	-2,517,764
offset	-1,796,832	-1,793,246	-1,770,809	-2,801,919	-2,800,669	-2,792,627
variance	-1,921,647	-1,918,061	-1,895,624	-2,980,246	-2,978,996	-2,970,954

Table 4.8: The deviances and AIC/BIC statistics for the four models considered in this chapter.

It should be noted that there are other hidden parameters in the multilevel models considered in this chapter, for example, the proportionality constant ξ which models the fact that the peak width increases with the *m/z* value and the peak locations μ_j. However, the model deviances are so large compared with the number of hidden parameters that even if they were included the overall conclusions would remain the same.

When comparing the *m/z* values obtained from this analysis which exhibit different intensities between the groups with those from the MCMC analysis presented in chapter 3 we see that there are great similarities. The majority of the peak locations still indicate that the peak heights are lower in stage IV than in stage I. For both datasets the locations exhibiting significant differences between groups remain similar across the two analyses in chapter 3 and this chapter. In the breast cancer dataset the numbers of significant locations for the control/control comparisons are 40 and 45 for chapters 3 and 4 respectively. For the control/treated comparisons the corresponding number of significant locations is 30 in both chapters. For the melanoma dataset the numbers of significant locations remain similar between the two analyses at 40 and 41 respectively. However, for the breast cancer dataset there are a greater number of differences identified between the *adcon* and *adtax* groups in the multilevel analysis than in the MCMC analysis. This may be attributable to the omission of day information in the multilevel analysis.

The results obtained by Dryden et al. (2005) and Mian et al (2005) are also found in the multilevel analysis. In Dryden et al. (2005), however, no significant results are found at *m/z* values greater than 15,000 Da. In this chapter there are numerous examples. The number of significant results obtained may have been affected by small sample sizes and if more data was available the number may be reduced. Mian et al (2005) identified the *m/z* value 11,701 Da as a location where the two groups have highly different variability. This location is identified in the multilevel analysis as one where the two groups differ.

Chapter 5

Conclusions and Further Work

5.1 Conclusions

The two main objectives of this thesis were to model the spectra obtained from SELDI-TOF mass spectrometric analysis of cancerous cells and sera and to identify where significant differences in protein expression levels occurred between groups. This has involved the use of existing statistical techniques along with the development of a procedure to identify peaks present in the data.

The data obtained from mass spectrometry of proteins is high-dimensional in nature and thus the first task was to reduce the dimensionality of the problem. This was achieved in chapter 2 with the development of a peak finding algorithm. The algorithm worked by identifying the largest peak in the mean spectrum and fitting a Gaussian peak to this location with heights that differ between spectra to match the data. The effects of this peak are then subtracted from the data and the process repeated on the remaining data until a specified number of peaks have been fitted. By using this algorithm to identify peaks in the data we reduced the number of m/z values needed to describe a spectrum from $\approx 14,000$ to 150. Despite this 100 fold reduction, the ability to classify test spectra given a set of training spectra did not fall greatly. When the complete spectra were used, correct classification rates of 89% and 86% were obtained when using principal components analysis combined with linear discriminant analysis for the breast cancer and melanoma datasets respectively. After dimension reduction using the peak finding algorithm instead of principal components analysis, the respective correct classification rates were 84% and 81%. The maximal correct classifications when using the algorithm each used around 30 peak locations whereas information from all 14,000 m/z values was used to obtain the higher rates. Classification using peak locations provides us with much more interpretable reasons for classifying a particular spectrum into a group than when using the principal components. We can identify a particular m/z value which classifies well and then molecules around this m/z value can be further investigated by chemists to determine if they hold any potential information for drug development.

Methods were also developed in chapters 3 and 4 to determine where significant differences in protein expression levels occurred between groups. In chapter 3 the use of Markov chain Monte Carlo methods enabled a parametric model to be fitted to the data. The model parameters initially included peak locations, peak heights, a proportionality constant used to calculate the peak variances and a residual variance/precision parameter. The parameters were all updated using Metropolis-Hastings steps with normal proposal distributions with the exception of the residual variance parameter which was updated using a Gibbs sampling step. The likelihood expression was complex and the full conditional distribution could only easily be written down for the residual variance parameter. An adapting stage was built into the MCMC algorithm to fine tune the proposal variances so that the proposals were accepted between 40% and 60% of the time. The peak finding procedure from chapter 2 was used at the beginning of the MCMC algorithm to provide a suitable starting point for each of the chains. This helped to ensure that prominent peaks in the data were not missed by a bad choice of starting value. The model was extended via three more related models to include a combination of two offset peaks with different variances at each peak location. This model provided a better fit to the data than the initial model and a selection of *m/z* values were identified which exhibit differences between groups. The *m/z* values identified by this MCMC analysis showed similarities to those obtained by Dryden et al. (2005). It was also shown in chapter 3 that by using parallel processing on the GRID computing system we can greatly reduce the computational time needed to carry out MCMC analysis of the data.

An algorithm was developed to create a partition of the data so that each section could be modelled separately. The split points were placed at *m/z* values with low average intensity in an attempt to not split peaks across section. Using this algorithm on the two datasets resulted in 17 sections for the breast cancer and 19 for the melanoma. This partition was used for both the MCMC work and the multilevel modelling work to reduce computation time. The methods used in chapters 2 and 3 do not take into account the complete structure of the data. To incorporate this hierarchical structure in the data a multilevel modelling framework was used in chapter 4. Using the IGLS algorithm to estimate the parameter values a mean spectrum for each group was identified and differences in these mean spectra were calculated using approximate Z tests. The four models used for the Bayesian analysis were also considered here and the most appropriate model was determined by likelihood ratio tests and AIC statistics. In agreement with the Bayesian analysis it was found that the model consisting of offset peaks with differing variances was considered the most appropriate.

In summary, the overall results obtained from this research are shown in tables 5.1 and 5.2. These tables show the *m/z* values which have been identified as being important in differentiating between different groups. Only the most significant result for each *m/z* value is shown. At 5,416 Da both the *mcc/mct* and *adc/adt* comparisons are significant in the MCMC and multilevel analyses but the comparison with the largest statistic has changed between chapters.

Similar reasoning applies to the location at 8,110 Da.

***m/z* value**	**4,393**	**4,881**	**5,416**	**7,453**	**8,110**	**10,230**
% times one of top 6 best classifiers (chapter 2)	31.9	19.1	39.3	61.0	83.2	24.6
largest t statistic between groups (chapter 3)	-8.576 *adc/mcc* (3)	-8.803 *tdc/mcc* (3)	8.713 *mcc/mct* (4)	12.852 *adc/tdc* (3)	9.219 *mcc/mct* (4)	5.408 *tdc/mcc* (2)
largest Z statistic between groups (chapter 4)	14.800 *adc/mcc*	11.318 *tdc/mcc*	6.068 *adc/tdc*	11.290 *adc/tdc*	9.891 *adc/mcc*	13.102 *tdc/mcc*

Table 5.1: *m/z* values identified as important in the analyses between groups in the breast cancer dataset.

***m/z* value**	**2,495**	**2,771**	**3,316**	**3,885**	**8,949**	**28,160**
% times one of top 6 best classifiers (chapter 2)	15.1	16.1	17.3	97.1	16.5	31.0
t statistic between stages (chapter 3)	7.885	6.471	4.171	9.085	4.845	7.566
Z statistic between stages (chapter 4)	8.333	6.333	5.889	8.833	4.750	7.543

Table 5.2: *m/z* values identified as important between stages in the melanoma dataset.

These *m/z* values should be investigated more thoroughly to identify precisely which molecules are present at these locations. These molecules could then be studied to ascertain if they show any promise for drug development.

The work in this thesis has considered a variety of ways to analyse mass spectrometry data. There remain, however, many ways in which the methods currently in use could be improved. Some potential improvements are discussed in the next section.

5.2 Further Work

5.2.1 Different distributions for peaks

The model that has been developed in this thesis for proteomic spectra consists of a mixture of Gaussian peaks at differing locations along a spectrum. In chapter 3 it was noted that some of the peaks in the data do not appear to be Gaussian and that a large number of peaks appear to have longer right hand tails. This is a more obvious problem with the melanoma data although it does sometimes occur in the breast cancer data. The work presented in chapter 3 attempted to reduce this problem by considering the modelled peaks as a combination of two offset peaks with differing variances. This solution allowed the modelling of the longer right hand tails although, in some cases, the peak heights did not match the data exactly.

Another possible way to model the data would be to use other distributions for the peaks. Three possibilities are the lognormal, beta or gamma distributions and an example of each is shown in figure 5.1. When the parameter values are changed then the peaks can appear more symmetrical. This could

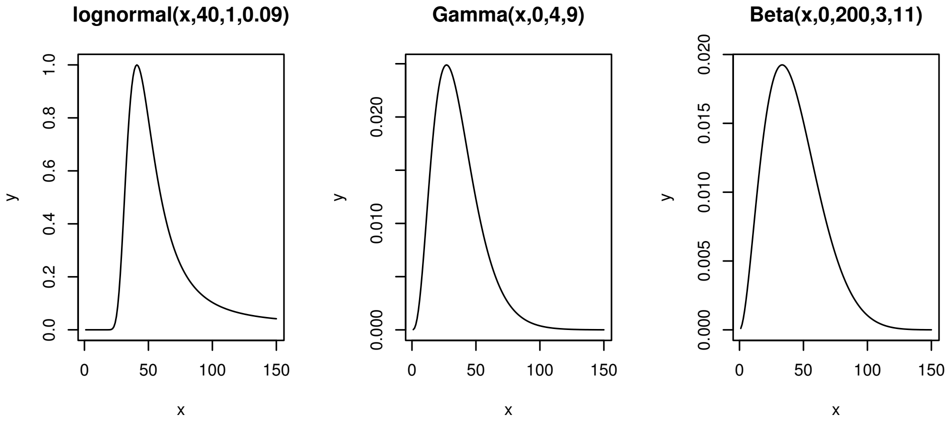

Figure 5.1: Plots of the lognormal, gamma and beta distributions showing how certain parameter values could be chosen to model non-Gaussian peaks.

prove useful in a dataset such as the breast cancer dataset studied in this thesis where there are a combination of symmetric and non-symmetric peaks. In the MCMC algorithm the parameters describing the distributions at each peak could be updated to accommodate this difference. Beta distributions have been used in this context by Müller et al. (2006) to model MALDI-TOF data although the number of observations in the dataset is small (17 control, 24 tumour). In their analysis the data are interpreted as a histogram and a mixture model is fitted. However, given the large number of 'observations' in the histogram the results of the analysis should not differ much from those already presented in this thesis.

5.2.2 Aligning Spectra

In the peak finding algorithm described in chapter 2 one of the main restrictions is that the peaks have common locations across all spectra. This is an important restriction which aids the interpretability of the results and allows comparison. However, it does sometimes result in slightly incorrect peak heights in the cases where the peak location is not an exact match for that particular spectrum. In these cases the modelled peak is matching to a point on the slope of a peak.

If we consider figure 5.2 we can see that this misalignment problem is quite common. It should also be noted that the misalignments for each of the two spectra shown in figure 5.2 are in the same direction for all the peaks in the range (and others not shown). This suggests that there has been a shift in the spectrum which has moved it a number of m/z values in a particular direction. The true peaks in figure 5.2 probably lie somewhere between the two examples shown.

To reduce this problem of shifted spectra we could use an algorithm to align all of the spectra before analysis starts. One possible method is that of dynamic programming as used by Glasbey et al. (2005). Specialised software is also becoming available for spectrum alignment, e.g. SpecAlign (Wong et al., 2005).

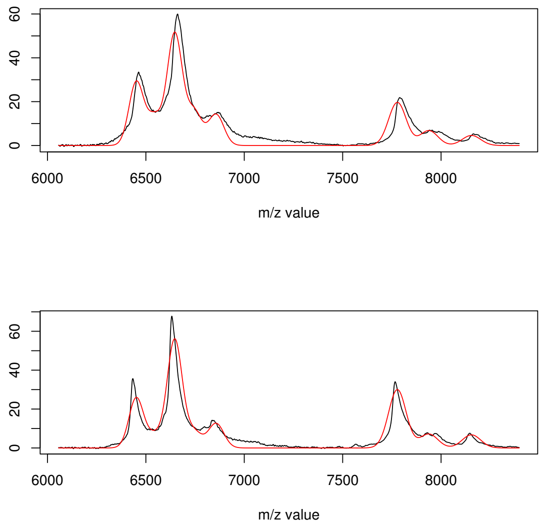

Figure 5.2: Plot of a section of the melanoma data and the fitted peaks obtained from the algorithm described in chapter 2. Note that the peaks around m/z values of 6,450, 6,700 and 7,800 Daltons are misaligned compared to the original data and this misalignment can occur in both directions.

5.2.3 Further Modelling

As explained at the beginning of chapter 4 it is important to take into account the complete structure of a dataset if it is known. This was considered in this thesis by using multilevel modelling to incorporate the hierarchical structure of the data. Each of the m/z values belongs to a particular spectrum and will be more similar to m/z values nearby than it will to others from another spectrum.

This worked well for the melanoma dataset as the structure of the data is relatively simple and conformed to the two level model explained in section 1.6.3. However, for the breast cancer dataset, the structure is not quite so simple. For this dataset not only do we have a *m/z* values within spectra hierarchy but there is also information concerning the day the spectrum was obtained and the experiment number. In this case we could include some peak/day interactions. By not taking into account the day information when analysing the breast cancer dataset in chapter 4 it is likely that we will have obtained an incorrect number of significant results.

Day information for the breast cancer dataset should also be incorporated into the MCMC work in chapter 3. As the model currently stands, the fact that the same samples are studied over a period of four days is ignored. To remedy this a piecewise linear function of time could also be fitted as part of the MCMC procedure.

Also, further modelling could be carried out in the MLwiN package by considering Bayesian fitting procedures for multilevel models. This would enable closer comparisons with the Bayesian approach of chapter 3

5.2.4 Clustering within groups

The ability to classify unknown spectra into the correct groups is important and could potentially enable physicians to determine the course of treatment for particular patients. In chapter 2 we achieved a rate of above 80% for correct classifications of test spectra for both the breast cancer and the melanoma datasets. However, depending on which spectra were used as training data, some spectra were misclassified.

The melanoma dataset identifies spectra as belonging to one of two groups - stage I or stage IV. From looking at the data it has been noted that there appears to be more than one type of spectrum present in stage IV. At an *m/z* value around 11,700 Da about 25% of the stage IV spectra are very different to the rest of the spectra in that group.

The presence of different clusters within a particular tumour classification could be important in identifying certain medical characteristics. For example, one cluster within the whole group could signify people in which the disease is more likely to recur in the future. The ability to predict recurrence would be of great benefit to patients and physicians alike.

5.2.5 Reversible Jump

The breast cancer and melanoma datasets were studied in chapter 3 using a Bayesian approach which utilised MCMC algorithms. This approach modelled the spectra as a series of peaks with differing heights and locations. However, one limitation of the method used in chapter 3 is that the number of peaks to be found in the data is fixed before any analysis starts. The peaks are identified using the peak finding algorithm described in chapter 2 which picks peaks in order of decreasing size. If too few peaks are fitted then this could result in some small but important peaks being ignored. In the opposite case, too many peaks could be identified which would cause overfitting to the data.

To remove this dependence on a user-specified number of modelled peaks, the MCMC algorithm could be adapted to use *reversible jump* methods. Reversible jump MCMC was introduced by Green (1995) as a method to simulate the posterior distribution when the number of parameters varies. By using this method, the initial set of peak location parameters can be increased or decreased at each update by the proposal of a *birth* or a *death* step.

For the algorithms used in chapter 3 this would mean that the number of peaks in the model could become a parameter itself and would not have to remain restricted to a fixed value - 150 in the current analysis.

References

1. Adam, B.L., Vlahou, A., Semmes, O.J. and Wright Jr, G.L. (2001), Proteomic approaches to biomarker discovery in prostate and bladder cancers, *Proteomics*, **1**(10), pp 1264-1270.

2. Adam, B.L., Qu, Y., Davis, J.W., Ward, M.D., Clements, M.A., Cazares, L.H., Semmes, O.J., Schellhammer, P.F., Yasui, Y., Feng, Z. and Wright Jr, G.L. (2002), Serum protein fingerprinting coupled with a pattern matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men, *Cancer Research*, **62**(13), pp 3609-3614.

3. Aitken, A. C. (1935), On least squares and the linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, **55**, pp 42-48.

4. Akaike, H. (1974), A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**(6), pp 716-723.

5. Baggerly, K.A., Morris, J.S., Coombes, K.R. (2004), Reproducibility of SELDI-TOF protein patterns in serum: Comparing data sets from different experiments, *Bioinformatics*, **20**(5), pp 777-785.

6. Ball, G., Mian, S., Holding, F., Allibone, R.O., Lowe, J., Ali, S., Li, G., McArdle, S., Ellis, I.O., Creaser, C. and Rees, R.C. (2002), An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers, *Bioinformatics*, **18**(3), pp 395-404.

7. Banks, R.E., Stanley, A.J., Cairns, D.A., Barrett, J.H., Clarke, P., Thompson, D. and Selby, P.J. (2005), Influences of blood sample processing on low-molecular-weight proteome identified by surface-enhanced laser desorption/ionization mass spectrometry, *Clinical Chemistry*, **51**(9), pp 1637-1649.

8. Benjamini, Y. and Hochberg, Y. (1995), Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **57**(1), pp 289-300.

9. Benjamini, Y. and Hochberg, Y. (2000), On adaptive control of the false discovery rate in multiple testing with independent statistics, *Journal of Educational and Behavioral Statistics*, **25**(1), pp 60-83.

10. Bernardo, J.M. and Smith, A.F.M. (1994), *Bayesian Theory*, Wiley, New York.

11. Boser, B.E., Guyon, I.M. and Vapnik, V.N. (1992), A training algorithm for optimal margin classifiers, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp 144-152.

12. Browne, W.J. and Draper D. (2000), Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models, *Computational Statistics*, **15**(3), pp 391-420.

13. Chen, G., Gharib, T.G., Huang, C.C., Thomas, D.G., Shedden, K.A., Taylor, J.M.G., Kardia, S.L.R., Misek, D.E., Giordano, T.J., Iannettoni, T.D., Orringer, M.B., Hanash, S.M. and Beer, D.G. (2002), Proteomic analysis of lung adenocarcinoma: identification of a highly expressed set of proteins in tumours, *Clinical Cancer Research*, **8**(7), pp 2298-2305.

14. Cortes, C. and Vapnik, V. (1995), Support-vector networks, *Machine Learning*, **20**(3), pp 273-297.

15. Diamandis, E.P. (2004), Mass spectrometry as a diagnostic and a cancer biomarker doscovery tool, *Molecular and Cellular Proteomics*, **3**(4), pp 367-378.

16. Dryden, I.L., Mian, S., Browne, W.J., Handley, K., di Nisio, R. and Rees, R. (2005), Statistical analysis of SELDI protein chip data from breast cancer cell lines exposed to chemotherapeutic agents, *In proceedings of LASR 2005: Quantitative Biology, Shape Analysis and Wavelets*, editors: Barber, S., Baxter, P.D., Mardia, K.V. and Walls, R.E., University of Leeds, pp 43-46.

17. Edwards, J.J., Anderson, N.G., Tollaksen, S.L., von Eschenbach, A.C. and Guevara Jr, J. (1982), Proteins of human urine II. Identification by two-dimensional electrophoresis of a new candidate marker for prostatic cancer, *Clinical Chemistry*, **28**(1), pp 160-163.

18. Fisher, R.A. (1936), The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, **7**, pp 179-188.

19. Gelfand, A.E. and Smith, A.F.M. (1990), Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**(410), pp 398-409.

20. Gelman, A., Roberts, G.O. and Gilks, W.R. (1996), Efficient Metropolis Jumping Rules, in *Bayesian Statistics*, **5**, pp 599-608, Oxford University Press.

21. Geman, S. and Geman, D. (1984), Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**(6), pp 721-741.

22. Gilks, W., Richardson, S. and Spiegelhalter, D. (1996), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London.

23. Glasbey, C.A., Vali, L. and Gustafsson, J.S. (2005), A statistical model for unwarping of 1-D electrophoresis gels, *Electrophoresis*, **26**(22), pp 4237-4242.

24. Goldstein, H. (1986), Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares, *Biometrika*, **73**(1), pp 43-56.

25. Goldstein, H. and Spiegelhalter, D.J. (1996), League Tables and their Limitations: Statistical Issues in Comparisons of Institutional Performance, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **159**(3), pp 385-443.

26. Green, P. (1995), Reversible jump markov chain monte carlo computation and bayesian model determination, *Biometrika*, **82**(4), pp 711-732.

27. Handley, K., Browne, W.J. and Dryden, I.L. (2005), Bayesian analysis of SELDI-TOF data, *In proceedings of LASR 2005: Quantitative Biology, Shape Analysis and Wavelets*, editors: Barber, S., Baxter, P.D., Mardia, K.V. and Walls, R.E., University of Leeds, pp 138-141.

28. Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York.

29. Hastings, W.K. (1970), Monte Carlo Sampling using Markov chains and their applications, *Biometrika*, **57**(1), pp 97-109.

30. Hortin, G.L. (2006), The MALDI-TOF mass spectrometric view of the plasma proteome and peptidome, *Clinical Chemistry*, **52**(7), pp 1223-1237.

31. Hotelling, H. (1931), The generalization of student's ratio, *Annals of Mathematical Statistics*, **2**(3), pp 360-378.

32. Hutchens, T.W. and Yip, T.T. (1993), New desorption strategies for the mass spectrometric analysis of micromolecules, *Rapid Communications in Mass Spectrometry*, **7**(7), pp 576-580.

33. Hyvärinen, A., Karhunen, J. and Oja, E. (2001), *Independent Components Analysis*, John Wiley & Sons, New York.

34. Izmirlian, G. (2004), Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial, *Annals of the New York Academy of Sciences*, **1020**(1), pp 154-174.

35. Jeffreys, H. (1946), An invariant form for the prior probability in estimation problems, *Proceedings of the Royal Society of London: Series A (Mathematical and Physical Sciences)*, **186**(1007), pp 453-461.

36. Kelly, F.P. and Ripley, B.D. (1976), A note on Strauss's model for clustering, *Biometrika*, **63**(2), pp 357-360.

37. Kirkwood, J.M., Atkins, M.B., Ernstoff, M.S., Gershenwald, J.E., Halpern, A.C., Thompson, J.A. and Weinstock, M.A. (2003),
http://www.melanomacenter.org/glossary/m.html.
Accessed 2007/3/5

38. Klose, J. (1975), Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals, *Humangenetik*, **26**(3), pp 231-243.

39. Lachenbruch, P.A. (1982), *Discriminant analysis: in Encyclopedia of Statistical Sciences*, edited by S. Kotz and N.L. Johnson, John Wiley & Sons, New York.

40. Mardia K.V., Kent J.T. and Bibby J.M. (1979), *Multivariate Analysis*, Academic Press, London.

41. Markham, A. (2005), http://info.cancerresearchuk.org/cancerandresearch/cancers/breast/?a=5441 Accessed 2007/3/5.

42. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953), Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, **21**(6), pp 1087-1091.

43. Mian, S., Ball, G., Hornbuckle, J., Holding, F., Carmichael, J., Ellis, I., Ali, S., Li, G., McArdle, S., Creaser, C. and Rees, R.C. (2003), A prototype methodology combining surface-enhanced laser desorption/ionization protein chip technology and artificial neural network algorithms to predict the chemoresponsiveness of breast cancer cell lines exposed to Paclitaxel and Doxorubicin under in-vitro conditions, *Proteomics*, **3**(9), pp 1725-1737.

44. Mian, S., Ugurel, S., Parkinson, E., Schlenzka, I., Dryden, I., Lancashire, L., Ball, G., Creaser, C., Rees, R. and Schadendorf, D. (2005), Serum proteomic fingerprinting discriminates between clinical stages and predicts disease progression in melanoma patients, *Journal of Clinical Oncology*, **23**(22), pp 5088-5093.

45. Müller, P., Do, K.A., Baggerly, K. and Bandyopadhyay, R. (2006), *Applied MCMC II: Protein Mass/Charge Spectra*, available at http://odin.mdacc.tmc.edu/∼pm/422/handout2.pdf.

46. Munro, N.P., Cairns, D.A., Clarke, P., Rogers, M., Stanley, A.J., Barrett, J.H., Harnden, P., Thompson, D., Eardley, I., Banks, R.E. and Knowles, M.A. (2006), Urinary biomarker profiling in transitional cell carcinoma, *International Journal of Cancer*, **119**(11), pp 2642-2650.

47. Nomura, F., Tomonaga, T., Sogawa, K., Ohashi, T., Nezu, M., Sunaga, M., Kondo, N., Iyo, M., Shimada, H. and Ochiai, T. (2004), Identification of novel and downregulated biomarkers for alcoholism by surface en-

hanced laser desorption/ionization-mass spectrometry, *Proteomics*, **4**(4), pp 1187-1194.

48. O'Farrell, P.H. (1975), High-resolution two dimensional gel electrophoresis of proteins, *Journal of Biological Chemistry*, **250**(10), pp 4007-4021.

49. Petricoin III, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C. and Liotta, L.A. (2002), Use of proteomic patterns in serum to identify ovarian cancer, *The Lancet*, **359**(9306), pp 572-577.

50. Poon, T.C., Yip, T.T., Chan, A.T., Yip, C., Yip, V., Mok, T.S., Lee, C.C., Leung, T.W., Ho, S.K. and Johnson, P.J. (2003), Comprehensive proteomic profiling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes, *Clinical Chemistry*, **49**(5), pp 752-760.

51. Qu, Y., Adam, B.L., Yasui, Y., Ward, M.D., Cazares, L.H., Schellhammer, P.F., Feng, Z., Semmes, O.J. and Wright Jr, G.L. (2002), Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from non-cancer patients, *Clinical Chemistry*, **48**(10), pp 1835-1843.

52. R Development Core Team (2005), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna.

53. Rasbash, J., Browne, W.J., Goldstein, H., Yang, M. et al. (2000), *A User's Guide to MLwiN* (Second Edition), Institute of Education, London.

54. Robbins, R.J., Villanueva, J., and Tempst, P. (2005), Distilling cancer biomarkers from the serum peptidome: high technology reading of tea leaves or an insight to clinical systems biology?, *Journal of Clinical Oncology*, **23**(22), pp 4835-4837.

55. Schwartz, G. (1978), Estimating the dimension of a model, *Annals of Statistics*, **6**(2), pp 461-464.

56. Sobin, L.H. and Wittekind, C. (2002), *TNM Classification of Malignant Tumours*, Wiley-Liss, New York.

57. Somorjai, R.L., Dolenko, B. and Baumgartner, R. (2003), Class prediction and discovery using gene microarray and proteomics mass spectrometry data: curses, caveats, cautions, *Bioinformatics*, **19**(12), pp 1484-1491.

58. Sorace, J.M. and Zhan, M. (2003), A data review and re-assessment of ovarian cancer serum proteomic profiling, *BMC Bioinformatics*, **4**(24), pp 1-13.

59. Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1996), *BUGS: Bayesian Inference using Gibbs Sampling, version 1.4.1*, MRC Biostatistics Unit, Cambridge.

60. Srinivas, P.R., Srivastava, S., Hanash, S. and Wright Jr, G.L. (2001), Proteomics in early detection of cancer, *Clinical Chemistry*, **47**(10), pp 1901-1911.

61. Stroustrup, B. (2003), C++ programming language, Bell Laboratories, New Jersey, USA.

62. Student (1908), The probable error of a mean, *Biometrika*, **6**(1), pp 1-25.

63. Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A. and Le, Q.T. (2004), Sample classification from protein mass spectrometry by 'peak probability contrasts', *Bioinformatics*, **20**(17), pp 3034-3044.

64. Uchida, T., Fukawa, A., Uchida, M., Fujita, K. and Saito, K. (2002), Application of a novel protein biochip technology for detection and iden-

tification of rheumatoid arthritis biomarkers in synovial fluid, *Journal of Proteome Research*, **1**(6), pp 495-499.

65. Vlahou, A., Schellhammer, P.F., Mendrinos, S., Patel, K., Kondylis, F.I., Gong, L., Nasim, S. and Wright Jr, G.L. (2001), Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine, *American Journal of Pathology*, **158**(4), pp 1491-1502.

66. Wani, M.C., Taylor, H.L., Wall, M.E., Coggon, P. and McPhail, A.T. (1971), Plant antitumor agents. VI. Isolation and structure of taxol, a novel antileukemic and antitumor agent from Taxus brevifolia, *Journal of the American Chemical Society*, **93**(9), pp 2325-2327.

67. Wiley J. & Sons (2006),
http://www.spectroscopynow.com/ftp_images/msprim_spark_f1c.jpg.
Accessed 2007/3/5.

68. Winston, W.L. (1995), *Introduction to Mathematical Programming: Applications and Algorithms*, Wadsworth Publishing Company, California.

69. Wong, J.W.H., Cagney, G. and Cartwright, H.M. (2005), SpecAlign - processing and alignment of mass spectra datasets, *Bioinformatics*, **21**(9), pp 2088-2090.

70. Xiao, Z., Adam, B.L., Cazares, L.H., Clements, M.A., Davis, J.W., Schellhammer, P.F., Dalmasso, E.A. and Wright Jr, G.L. (2001), Quantitation of serum prostate-specific membrane antigen by a novel protein biochip inmmunoassay discriminates benign from malignant prostate disease, *Cancer Research*, **61**(16), pp 6029-6033.

71. Yanagisawa, K., Shyr, Y., Xu, B.J., Massion, P.P., Larsen, P.H., White, B.C., Roberts, J.R., Edgerton, M., Gonzalez, A., Nadaf, S., Moore, J.H., Caprioli, R.M. and Carbone, D.P. (2003), Proteomic patterns of tumour subsets in non-small-cell lung cancer, *The Lancet*, **362**(9382), pp 433-439.

72. Yasui, Y., Pepe, M., Thompson, M.L., Adam, B.L., Wright Jr, G.L., Qu, Y., Potter, J.D., Winget, M., Thornquist, M. and Feng, Z. (2003), A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection, *Biostatistics*, **4**(3), pp 449-463.

73. Zhang, Z., Barnhill, S.D., Zhang, H., Xu, F., Yu, Y., Jacobs, I., Woolas, R.P., Berchuck, A., Madyastha, K.R. and Bast Jr, R.C. (1999), Combination of multiple serum markers using an artificial neural network to improve specificity in discriminating malignant from benign pelvic masses, *Gynecologic Oncology*, **73**(1), pp 56-61.