Pattern Recognition Analysis of *In Vivo* Magnetic Resonance Spectra

Anne Rosemary Tate

432

September, 1996

ISSN 1350–3162

UNIVERSITY OF

SUSSEX

AT BRIGHTON

Cognitive Science

Research Papers

Acknowledgements

Firstly I would like to thank my supervisor Des Watson for providing advice and support throughout the three years of this study and for always 'being there'. Secondly I would like to thank the members of my DPhil committee: Peter Williams and David Young for their invaluable advice and support. Thirdly I would like to thank a number of people at the Robert Steiner unit at Hammersmith Hospital for their help. These include David Bryant, Jimmy Bell, Louise Thomas, Nadeem Saeed, Glyn Coutts and Simon Taylor-Robinson. Fourthly I would like to thank the team at St George's Hospital, especially John Griffiths, Loreta Rodrigues and Roy Mazucco.

I would also like to thank the excellent support team and all my friends in COGS who have helped me considerably. These include Jo Brooks, Theo Arvanitis, Joe Wood, Jim Stone, Julian Budd, Magdalena Portmann, Roger Sinnhuber, James Goodlet, Julie Coultas, Sharon Groves and Paul Hackney. In particular I would like to thank Stephen Eglen who has given me a great deal of help and encouragement. I would also like to thank Joanne Mathias, Paul Tofts, Wael El-Deredy, David Hitchin, Yvette Mallet, Murray Alexander and many others too numerous to mention.

Lastly I thank my long-suffering daughters, Catherine and Lisa, for putting up with my long periods of non-communication and my obsession with wavelets.

Pattern Recognition Analysis of *In Vivo* Magnetic Resonance Spectra

Anne Rosemary Tate

Submitted for the degree of D. Phil.

University of Sussex

September, 1996

Abstract

Magnetic resonance spectroscopy (MRS) provides a unique non-invasive method for obtaining information on the biochemistry of living tissue *in situ*, and therefore has great potential as a clinical tool. However, presently *in vivo* MRS is used mainly for research, rather than for clinical applications.

There are a number of reasons for this. The information may be difficult to extract from the spectrum due to low signal-to-noise ratio and other problems associated with obtaining a signal from living tissue. Interpretation may be difficult due to the large number of metabolites represented by the spectra. Another problem is that most current methods for analysing MRS data are targeted at providing information on specific metabolites, rather than the more general information appropriate for clinical applications, such as the disease stage or state of the tissue being examined.

This thesis shows how pattern recognition techniques may be used to help overcome these problems and to provide methods for classifying *in vivo* spectra according to their tissue type. A prototype system for classifying spectra is developed using features that are extracted automatically, using the whole spectrum, rather than selected peaks. These features were selected purely on the basis of their power to discriminate between different types of spectra, using no prior knowledge of biochemistry. Among the techniques used were wavelets, principal component analysis and linear discriminant function analysis. These techniques were tested on two sets of *in vivo* data: 75 ^{13}C spectra obtained from healthy human volunteers from three different dietary groups of adipose tissue in the leg and 55 ^{31}P spectra obtained from tumorous and normal tissue in rats. For both datasets most of the spectra were assigned to their correct groups (94% of the ^{13}C and 86 – 100% of the ^{31}P spectra) without the need for explicit identification or measurement of peaks.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Nuclear Magnetic Resonance in Medicine

The use of Nuclear Magnetic Resonance (NMR) in medicine allows us to 'see' what is going on inside the body without carrying out invasive surgery or inserting optical instruments. NMR is not unique in this; there are other techniques for imaging the body such as X-rays and ultrasound. However, unlike other methods, NMR makes it possible not only to visualise anatomical structure with magnetic resonance imaging (MRI), but also to investigate physiological function with magnetic resonance spectroscopy (MRS). The extra dimension of information offered by magnetic resonance spectroscopy and also the fact that the technique has no known harmful effects makes NMR a unique and powerful imaging technique for clinical medicine.

NMR is concerned with the behaviour of atomic nuclei and their interaction with electromagnetic radiation. Certain nuclei, for example those of hydrogen (^{1}H), carbon (^{13}C) and phosphorus (^{31}P) 'resonate' when exposed to electromagnetic radiation at a particular frequency. This frequency is dependent on the type of nucleus and also on the intensity of the surrounding magnetic field. An NMR signal is produced by inducing nuclei of interest to resonate by exposing them to a pulse of radiation at their resonance frequency, and then allowing the nuclei to relax when they will release radiation at this same frequency. Because the strength of the resulting signal will depend on the number of nuclei present, it can be used to give a measure of the proportion of nuclei in a sample.

MRI depends on the fact that it is possible not only to obtain a measure of the nuclei resonating within a sample but also to spatially encode this measure. MRI is normally based on the ^{1}H nucleus (proton) and uses the fact that living tissue is largely composed of water, which in turn contains protons. The relative number of protons in different locations in a sample can be deduced from the NMR signal and displayed as an image. MRI was developed in the 1970's and has recently progressed to being a major imaging modality in clinical medicine. It has the advantage over the other most common form of imaging, X-ray, in that it does not, as far as is known, cause any harm to the patient. It has proved particularly successful in visualising organs such as the brain, where it can often eliminate the need for investigative surgery.

MRS is based on the fact that a nucleus will resonate at a slightly different frequency depending on its molecular environment. This phenomenon, known as 'chemical shift', is due to the fact that atoms and molecules surrounding a nucleus produce a shielding effect which influences its

local magnetic field. The relative numbers of nuclei resonating in different molecular sites in a sample can be deduced from the NMR signal. The peaks in the NMR spectrum represent nuclei resonating at slightly different frequencies, and the quantities of certain substances can be calculated by measuring the area under each peak. MRS is widely used tool in analytical chemistry where it is used, for example to elucidate the chemical structures of molecules. However it is not limited to the chemical laboratory as it can also be used to study living tissue. Spectra can be obtained *in vivo* allowing detailed biochemical information to be obtained non-invasively from patients.

MRS provides a unique means of observing living biochemistry *in situ* and thus has great potential as a clinical tool. However, while it is now widely used as a research tool, its use in clinical medicine has been so far slow to develop. This is partly due to the high cost of acquiring *in vivo* data and also because difficulties associated with obtaining a signal from living tissues, especially at the low magnetic fields acceptable for human patients, may make accurate quantification of the metabolites represented by the spectra very difficult.

Another reason that MRS is not yet widely used in clinical medicine is that spectral analysis and interpretation is a very time-consuming task which requires considerable expertise by a highly trained operator. In recent years the advances in technology have resulted in a huge information explosion, but, while it has become much easier to collect and process data, the development of methods for extracting meaningful and useful information has lagged behind. MRS is a good example of this; extremely sophisticated techniques have been developed for collecting MRS data, but relatively few for automating the analysis and interpretation of this data once it has been obtained.

Although there are a number of computer-based methods available for analysing spectra, most of these still need considerable interaction by the user in order to determine which metabolites are present in the spectrum, and in which proportions. Once the metabolites have been quantified it is up to the user to apply an understanding of biochemistry to draw conclusions from the composition of the sample about the probable nature of the tissue. The time and expertise required for the currently available methods to analyse and interpret spectra *in vivo* have contributed to the fact that MRS has remained primarily a research tool, despite its great potential for clinical applications.

If MRS is to realise its full potential as a clinical tool it will be essential to have reliable automated methods for analysing and interpreting MRS data. In particular it will be necessary to have methods which are specifically targeted to providing the kind of information that is required for clinical use. Developing and investigating such methods is the aim of the work described in this thesis.

1.2 Analysis of Spectra for Clinical Applications

In order to develop methods specifically targeted to providing the kind of information that is required for clinical use, it is first necessary to consider what kind of information will be needed, and what questions will need answering. As in all areas of data analysis it is important to match the methods of analysis to the specific questions to be addressed, not only to obtain the right answer, but also because the use of inappropriate methods may result in losing potentially relevant information.

Most current methods for spectral analysis aim at explicit quantification of the individual metabolites in individual spectra. This quantification may be essential for ascertaining the biochemical structure of the tissue being examined, as will be the case when the data are to be used for research purposes. However explicit quantification of metabolites may not be necessary, or

indeed desirable, if the aim is to use MRS as an aid to the decision making process of clinicians, for example to help them make a diagnosis. In this case, more abstract information will be required about the tissue being examined, such as its disease type or stage. In order to do this it will be necessary to have methods which can confidently discriminate between spectra of different clinical types.

Distinguishing between spectra of different clinical types is essentially a problem of classification, i.e. the fitting of chemical sample analyses or patient tests into a well-defined class. Classification methods are usually described under the general heading of 'pattern recognition', a term which encompasses a wide range of techniques which try to find patterns in groups of data and to distinguish between different subgroups. Once these patterns are identified they can be used to assign unknown individuals to a particular classification.

Feature extraction is concerned with finding the best patterns to discriminate and classify the data. In the case of MRS spectra this means choosing appropriate measurements to represent the spectra, and then finding which combination or subset of these measurements provides the best discrimination.

1.3 The Pattern Recognition Approach

Pattern recognition (PR) is a discipline which is devoted to extracting relevant information from data by identifying meaningful patterns. PR methods can be used to attempt to automate tasks that are carried out naturally by the human sensory systems, such as vision or speech recognition. Alternatively, and more realistically, they can be used to help make sense of multidimensional numerical data which we find much more difficult to interpret. It is thus a very appropriate methodology for the interpretation of MRS data. It is particularly appropriate for the purpose of this thesis because:

- Most clinical applications, such as medical diagnosis or the study of a patient's response to drug treatments will require the classification of the spectra into distinct groups. Classifying data is one of the main objectives of PR analysis, in fact the terms classification, discrimination and pattern recognition are often used synonymously.
- PR is usually a computer-based approach, designed to handle large amounts of data at the same time and is thus ideal for developing an automated system.

There are many other advantages of the PR approach which will be discussed in later chapters.

It is convenient to divide the PR process into three stages, although these stages normally overlap considerably:

- acquisition and pre-processing of the data
- extraction of features which can be used to describe the data
- classification of the data.

A computer system for classifying spectra will need automated methods for each of these three stages. Since the object of this thesis is to develop methods for discriminating between spectra using no prior knowledge of biochemistry, the methods investigated here use the whole spectrum

and features are chosen solely on the basis of their power to discriminate between different classes of spectra.

The benefits of such an investigation are twofold: firstly the development of such techniques will be necessary if the spectra are to be used as a routine aid to the medical decision process. Secondly PR involves not only methods for classifying data but also finding the important discriminatory features. It can therefore provide an alternative 'view' of the data from that obtained by the more traditional methods of spectral analysis.

1.4 Thesis Structure

The structure of the thesis is as follows. In the next chapter on MR theory the phenomenon of MR and its potential applications in clinical medicine are explained in enough detail to give necessary background for the particular problems that will need to be addressed. Particular emphasis is placed on the potential relevance of the technique in clinical medicine and the difficulties that need to be surmounted in order that MRS can be a useful tool. The aim of the third chapter is to explain why and which PR methods were suitable for this project, and to give a brief history of their use in MRS data analysis. Since the main aim is to develop fully automated methods for feature extraction, most attention will be given to methods and previous studies which work towards this aim. The fourth chapter concentrates on the actual methods that were used in developing a prototype system for analysing *in vivo* data. In particular details are given of the methods used to process the data before statistical analysis and methods used to extract and select features. The methods explained in detail are: PCA, wavelet analysis, discriminant analysis and feature selection methods.

The fifth chapter presents the results using these methods on two data sets: ^{13}C spectra obtained *in vivo* from healthy human volunteers, of adipose tissue in the leg and ^{31}P spectra obtained *in vivo* from tumorous and normal tissue in rats.

The sixth chapter is devoted to discussion and future work.

Chapter 2

Nuclear Magnetic Resonance Theory

Nuclear Magnetic Resonance (NMR) is concerned with the paramagnetic behaviour of atomic nuclei and their interaction with electromagnetic radiation. Certain nuclei, such as the ^{1}H nucleus (the proton) and the ^{31}P nucleus, have nuclear spin – we can think of a nucleus as spinning around its own axis in the same way that the Earth turns around its axis, with an associated angular momentum. The spinning of these charged particles generates a magnetic moment along the axis of spin so that the nucleus can be regarded as a tiny bar magnet with its axis along the axis of rotation [Gadian, 1995]. If a sample containing such nuclei is placed in a strong magnetic field, some of these nuclear magnets will align themselves with the direction of the field, similar to the way compass needles point in the direction of the Earth's magnetic field. However, unlike a compass needle, the magnetic moment can have more than one orientation depending on its spin quantum number. For example the hydrogen nucleus (which has spin quantum number I = $\frac{1}{2}$) can have one of two orientations with respect to the applied field: parallel, in which the nuclear magnetic moment is aligned with the magnetic field, and anti-parallel when it is aligned against it. These two orientations have different energies associated with them, the parallel orientation having the lower energy. The nuclear magnets interact very weakly with the applied magnetic field, and therefore the values of the energy separation are low [Gadian, 1995].

Normally there will be a slightly higher proportion of nuclei in the lower energy state, but if a pulse of electromagnetic radiation at the correct frequency is applied to the sample, some of the nuclei in the lower energy state will absorb energy and 'flip over' into to the anti-parallel orientation resulting in a higher proportion of nuclei now being in the high energy state. When the pulse ceases, the nuclear magnets return to equilibrium and release this energy. This energy can be detected in a radio-frequency (RF) coil, inducing a voltage which is the MR signal. This signal provides qualitative and quantitative information about the nuclei which can be utilised to produce either an image or a spectrum. Images are normally based on the density of hydrogen nuclei in the body while spectra represent the densities of a certain nucleus in its different molecular environments.

For nuclei with spin numbers greater than $\frac{1}{2}$ there will be more than two possible orientations and in each case a set of equally spaced energy levels result. The radiation that is used, unlike other forms of medical imaging, is in the form of radio waves, which are at a much lower frequency, than X-rays (and hence have lower energy) and are thought to be harmless [Gadian, 1982].

The aim of this chapter is to introduce NMR at a level of abstraction appropriate to this project

and therefore only a very simplified description of the NMR process is given in the following sections. Full details of the physics of the NMR process can be found in numerous texts, for example [Hennel and Klinowski, 1993] [Slichter, 1989].

2.1 Nuclear Magnetic Resonance

The absorption of energy described above by the nuclear magnets is called nuclear resonance. Nuclei of certain atoms resonate when stimulated with radio waves of exactly the right frequency, called their Larmor frequency [Gadian, 1995]. Different nuclei resonate at different frequencies and then release electromagnetic radiation at this same frequency when they return to equilibrium. Images are usually based on ^{1}H nuclei. The most common nuclei used for medical applications of MRS are ^{1}H, ^{31}P, ^{13}C, ^{23}Na and ^{19}F.

Apart from the nature of the nucleus, the resonance frequency also depends on the strength of the surrounding magnetic field. The Larmor frequency for a particular nucleus is calculated using the following equation known as the Larmor equation:

$$\omega = \gamma B_{eff} \tag{2.1}$$

where ω is the Larmor angular velocity of precession, γ is the nucleus' gyromagnetic ratio, and B_{eff} is the effective magnetic field, that is the strength of the magnetic field surrounding the nucleus.

From this equation it can be seen that to stimulate a particular nucleus (and also to interpret the resulting signal) it is necessary to know not only its gyromagnetic ratio, which is a constant, but also the strength of the effective magnetic field B_{eff} [Gadian, 1995]. There are two factors which determine the strength of the magnetic field surrounding an individual nucleus. The main factor is the strength of the applied field B_0, that is the strength of the magnet that is being used. The second factor is the small local variations in field strength due to magnetic perturbations caused by surrounding electrons or other nuclei. This means that nuclei in different molecules, or located in various chemical groups in the same molecule will experience slightly different magnetic field strengths and will therefore resonate at a slightly different frequencies. The total effective field can be written as

$$B_{eff} = B_0(1 - \sigma) \tag{2.2}$$

where σ is the shielding or screening constant which expresses the contribution of the small secondary field generated by the electrons and which depends on the chemical environment of each nucleus [Gadian, 1995]. The difference in resonance frequency, caused by these local variations is called the chemical shift and is the basis for magnetic resonance spectroscopy (MRS).

The steps for NMR, both for imaging and spectroscopy are:

- Place the sample (e.g. the part of the body to be scanned) in a strong, homogeneous magnetic field.
- Excite the sample with a pulse of electromagnetic radiation to stimulate the nuclei to resonate. The frequency of the radiation used is determined by which nuclei we wish to observe, calculated using the Larmor equation.

- Allow the nuclei to 'relax' – that is return to equilibrium. As they do so they emit radiation at the same frequency at which it was absorbed and this can be picked up as a signal. This signal, called the free induction decay (FID) signal, provides the information to create an image or spectrum.

For MRS, the scan will cover a range of frequencies, since the resonance frequency for the nucleus of interest will vary according to its chemical environment. Rather than scan at each different frequency separately, it is usual to generate a RF pulse ranging over a selected bandwidth. Using Fourier transformation, the resulting FID signal, which is a function of time, can be analysed mathematically for the frequencies that it contains, to obtain a spectrum which is a function of frequency [Bracewell, 1986]. The intensity of the signal at a particular frequency is directly related to the number of the nuclei in the sample resonating at that frequency.

The intensity of the MR signals are dependent not only on the concentrations of the nuclei that give rise to the signals but also on many other parameters, including the time that the magnetisation of the sample takes to return to equilibrium (relaxation). This process is characterised by two relaxation times, T_1, the spin-lattice relaxation time and T_2, the spin-spin relaxation time. T_1 is the time constant for the recovery of magnetisation of the sample along the direction of B_0 and T_2 is the time constant for the recovery of magnetisation in the plane perpendicular to B_0. In general T_2 is significantly shorter than T_1[Sanders and Hunter, 1993].

2.2 The NMR Spectrum

The NMR spectrum represents the quantity of nuclei resonating at each frequency within the chosen bandwidth of radiation. The x-axis of the spectrum represents resonance frequency and the y-axis represents the intensity of the signal [Spisni, 1992]. The nuclei resonating at a particular frequency are normally quantified by measuring the areas under the 'peaks' at the various frequencies. Figure 2.1 shows an example of an FID signal and ^{13}C spectrum (the real part – see Section 2.3) obtained using a surface coil on a human thigh.

In order to make data collected on different instruments comparable, the resonance frequencies are always quoted in parts per million (ppm) from an arbitrarily chosen reference frequency. It is traditional to display the spectrum with the ppm scale decreasing along the x-axis. The left hand side of the spectrum is referred to as the low field region and the right hand side is called the high field region.

Fig 2.1 (a) shows the FID signal from which the spectrum was obtained. Fig 2.1 (b) shows the signal after Fourier transformation.

The ^{13}C spectrum shown above is an example of an spectrum obtained using a surface coil, which is based on the signal acquired from the area (i.e. the subcutaneous fat) directly underneath the coil. It is also possible to produce localised spectra from well-defined volume elements (voxels) at selected locations in the body using more sophisticated localisation techniques [Andrew *et al.*, 1990] [Cady, 1990].

Localisation techniques (other than surface coil techniques) are generally performed after selecting the region of interest using MRI. They fall into two categories: single voxel methods which involve obtaining spectra from a single volume of interest, and chemical shift (or spectroscopic) imaging methods in which spectra are acquired from multiple voxels. The advantage of localised spectroscopy is the ability to select the region of interest to be studied, for example a tumour or specific part of the brain. Chemical shift imaging has the additional advantage that it provides the

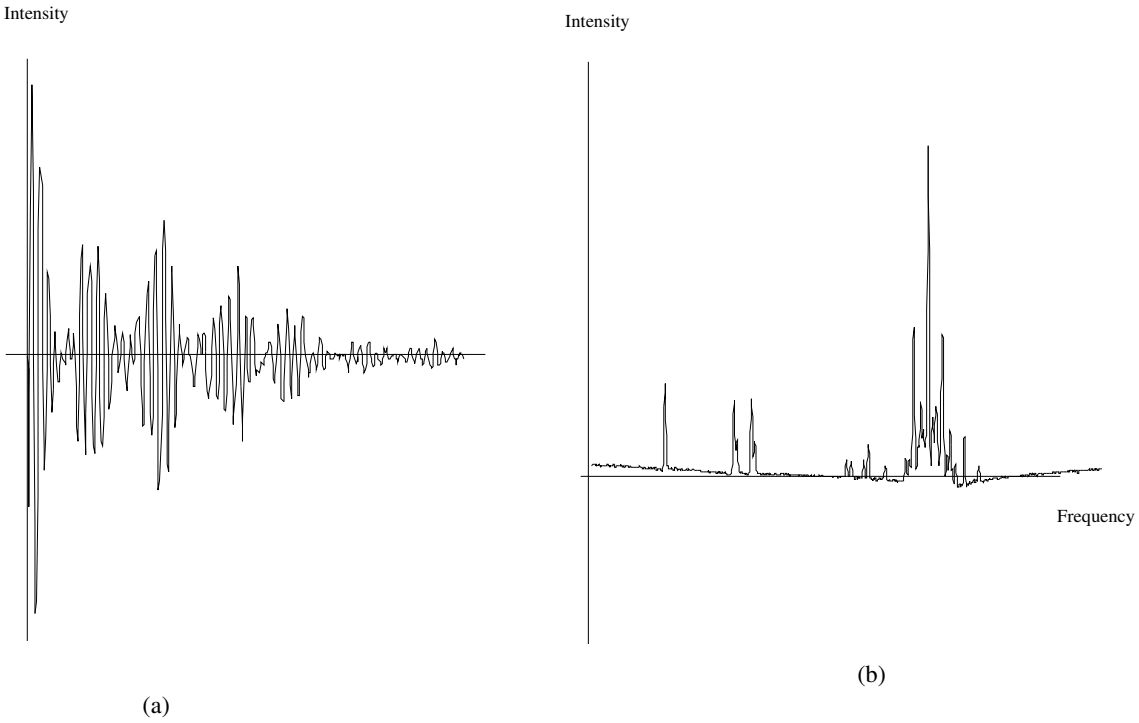

Figure 2.1. ^{13}C FID signal (a) and spectrum (b) from a human thigh

ability to acquire images that reflect the spatial distribution of metabolites. Localised spectra usually take longer to acquire than spectra obtained using a surface coil, and interpretation of spectra is generally more difficult, since the signal-to-noise ratio is lower and a necessary time delay in acquiring the signal may result in distortion of the shape of the spectrum [Wang, 1992].

It can be seen that some of the peaks in the above spectrum appear to be duplicated, for example the pairs of peaks in the group of four peaks towards the left-hand side of the spectrum look remarkably similar, and also the large group of peaks towards the right-hand-side appear to be symmetrical about the largest peak. This is due a phenomenon known as spin-spin coupling, where the spectral resonance arising from the same component is split into two or more components. This splitting is caused by an interaction between neighbouring nuclear spins which is transmitted by means of the electrons in the bonds joining the nuclei. Because the coupling effect may make the spectrum very complicated, a process known as proton decoupling is often used. This has the effect of collapsing multiplet signals into single lines. For more details of this effect see [Gadian, 1995].

2.3 Acquisition and Processing

In Fourier-transform NMR, the radio-frequency field that is used for excitation of signals is applied in the form of pulses. The bandwidth of these pulses, i.e. their effective spread in frequency, is sufficiently large to excite all of the nuclei within the required frequency range. The resulting signal, the FID, is a decaying signal which decays to zero with the time constant T_2^*. T_2^* differs from T_2 in that it incorporates the effect of field inhomogeneities as well as intrinsic relaxation effects. This signal is a superposition of the signals from resonant nuclei in the different chemical sites.

In general the spectral line widths are dictated by molecular mobility. Thus spectra reveal

narrow signals from metabolites which have a high degree of mobility whereas macromolecules which are highly immobilised produce very much broader signals, which are either invisible or appear as a broad hump underlying the signals from the metabolites. However linewidths and lineshapes can also be influenced by a range of other factors. Magnetic field homogeneity is a critical issue in NMR which affects the spectral linewidths and lineshapes, and therefore the resolution of the spectra. In order to obtain a high field homogeneity, the acquisition of the FID is preceded by a procedure known as 'shimming' which involves adjustments that are designed to optimise the field homogeneity for each given study.

Typically the signals that constitute a spectrum may have a centre frequency of, say, 63 MHz, occupying a frequency range of, say, a few kilohertz [Gadian, 1995]. Because the detection of signals at such high frequencies is difficult, the signal must be mixed with a reference frequency. The signal thus obtained is a much lower frequency difference signal. At such high frequencies, it is normal to subtract the centre frequency corresponding to the applied field. In order to determine the sign of the frequency of the signals compared with the reference frequency, the NMR receiver incorporates quadrature detection which results in two free induction decays 90° out of phase with each other being produced. See [Gadian, 1995] [Hennel and Klinowski, 1993] or [de Certaines *et al.*, 1992] for a full description of this technique.

Since the sensitivity of metabolites measured by MRS is low, the signal will usually be very weak in comparison with the random electrical noise in the system. To enhance the signal-to-noise ratio, a process called 'data accumulation' is carried out in which a large number of FIDs are acquired and added together [Andrew *et al.*, 1990] [Cady, 1990] [Gadian, 1982]. The accumulation of N consecutive acquisitions has the effect of increasing the signal-to-noise ratio by a factor of $\sqrt{N}$ because the signal increases by a factor of N whereas the background noise, being random, increases by $\sqrt{N}$.

Once the required number of signals have been acquired, Fourier transformation is carried out to produce a spectrum. In order to enhance the signal-to-noise ratio of the signal it is common at this stage to multiply the signal by a decaying exponential function. This process called apodisation ('cutting the feet off'), has the effect of reducing noise by preferentially lending more weight to the initial part of the free induction decay, where the signal-to-noise ratio is high, than to the latter part where the ratio is much lower [Gadian, 1982]. While the signal-to-noise ratio is considerably enhanced, this process does have the effect of increasing the linewidth, and therefore decreasing the spectral resolution. This is particularly undesirable when the peaks are close together or overlapping. The optimal signal-to-noise ratio in the spectrum is achieved when the decaying exponential has the same time constant as the free induction decay (T_2^*) [Gadian, 1995].

The result of the Fourier transformation is two components called the real and imaginary terms. Ideally the real part of the spectrum will be in 'absorption' mode and the imaginary part in 'dispersion' mode [Gadian, 1982]. The absorption mode spectrum is the form that characterises the absorption of energy by the nuclear spins and is the form that is normally displayed. Ideally the absorption mode has a Lorentzian lineshape $g(\nu)$ the equation for which is

$$g(\nu) \propto \frac{T_2}{1 + 4\pi^2 {T_2}^2 (\nu - \nu_0)^2} \tag{2.3}$$

where ν_0 is the resonance frequency. In practice the real part may not be a pure absorption mode spectrum, but may contain some dispersion mode components and vice versa. This is because the phase of the FID and its resulting Fourier transform depend on the settings of the NMR experiment. In order to produce a spectrum in pure absorption mode, it may be necessary to carry out phase correction after Fourier transformation [Gadian, 1995].

The phase correction makes use of both real and imaginary parts of a spectrum. A phase

change of θ corresponds to a manipulation of each data point according to the following equations:

$$r_2 = -i_1 sin\theta + r_1 cos\theta \tag{2.4}$$

$$i_2 = r_1 sin\theta + i_1 cos\theta \tag{2.5}$$

Where r_2 and i_2 correspond to the real and imaginary components of each point of the spectrum following phase correction and r_1 and i_1 correspond to the points prior to correction [Gadian, 1995]. The phase can be constant over the spectrum or it may depend linearly or polynomially on the offset frequency. Zeroth order correction, where a constant phase correction is applied, and first order correction, where a phase adjustment that is linearly related to the frequency is often necessary [Hennel and Klinowski, 1993].

2.4 Type of Information Available from *in vivo* Magnetic Resonance Spectroscopy

MRS provides information on the types of metabolites in the tissue. It also provides a means of measuring these metabolites. In addition to giving information about the concentrations of specific metabolites, MRS provides information about the intracellular environment of the metabolites. For example ^{31}P spectroscopy gives information about intracellular pH and the binding of Mg+ ions to ATP. In this, MRS has a significant advantage over other methods of chemical analysis when this information is normally lost because the tissue must be destroyed.

The nuclei which have the widest applicability for clinical MRS are ^{1}H, ^{13}C and ^{31}P. The ^{1}H nucleus is found in a vast number of biologically important molecules, and is the most sensitive stable nucleus. *In vivo* studies using the ^{1}H nucleus are handicapped, however, by the very narrow range of chemical shifts encountered (covering about 8 ppm, so that the magnetic field homogeneity is critical) and by the large number of metabolites which give many overlapping peaks and produce very complex spectra. In addition to the large number of resonances, many tissues have a high water content. While the existence of a large concentration of water in tissue is a very great advantage for MRI, it has the disadvantage for ^{1}H MRS in that weaker signals from other compounds within a sample can be overshadowed by the large peak representing water. Various methods have been developed to deal with this problem. These include methods of modifying the data acquisition process so that the influence of the water signal is reduced and also methods involving subsequent processing [Cady, 1990].

^{31}P spectra of living tissues and organs are comparatively simple compared with those of ^{1}H, as the number of metabolites containing this nucleus are fewer, and also because they occur in only a few molecular configurations, unlike the ^{1}H spectrum. However, this nucleus has proved one of the most important in the MRS investigation of living systems. This importance can be largely attributed to the presence of the nucleus in such metabolites as PCr, ATP and Pi, which are important in energy metabolism [Cady, 1990].

The ^{13}C isotope nucleus has a much lower sensitivity than those of either ^{1}H or ^{31}P. However, it has also proved useful for MRS studies of living tissues. One advantage of ^{13}C is the fact that it is possible to obtain ^{13}C enriched compounds which can be used for "tracer" studies of the chemical pathways utilised by a particular metabolite. Another advantage is that the chemical shift range in which resonances are observed (about 200 ppm) is large compared with ^{1}H spectra (8 ppm) and ^{31}P (40 ppm). For a review of *in vivo* ^{13}C spectroscopy in humans see [Beckmann, 1992].

For full details of the biochemical information available from MRS see, for example, [Andrew *et al.*, 1990] [Cady, 1990] [Gadian, 1982] [de Certaines *et al.*, 1992].

2.5 Clinical Applications of MR Spectroscopy

MRS was first developed in the 1940's and it has become a well established and important tool in analytical chemistry where it is used, for example, for elucidating chemical structures of compounds. The potential of MRS for biology was also appreciated around this time, but experiments were limited in scope by the relatively poor quality of the instrumentation that was then available. With the development of high-field superconducting magnets in the late 1960's together with the emergence of Fourier transform NMR, the scope of NMR was revolutionised and it became possible to use MRS to study proteins and other biological molecules. This led to the realisation that NMR might have extensive applications in the study of the metabolism of living systems [Gadian, 1995].

The clinical applications of MR spectroscopy have been slow to develop compared with MRI. This is partly because of the low signal-to-noise ratio, relatively poor spatial resolution and longer acquisition times caused by the low sensitivity of the technique, but also because MRS is more technically demanding than MRI and requires higher magnetic fields. Another reason is the fact the instrument manufacturers have been uncertain about the economic rewards of MRS and investment in this area has been much smaller than in MRI, resulting in slower development of tools for clinical investigation. This is particularly true for tools for classification and interpretation of spectra. Notwithstanding these problems, there is growing evidence that MRS provides unique clinical information which in some cases directly assists clinical diagnosis and choice of therapy. A brief résumé of some of the most promising applications is given below.

One of the main areas in which MRS shows great potential as a clinical tool is in the diagnosis and treatment of cancer. Both ^{31}P and ^{1}H spectra show different metabolite patterns according to the type of tumours [Howells *et al.*, 1992b] [Howells *et al.*, 1993a]. It has been shown that MRS can be successfully used to discriminate between different types of human brain tumours and between normal tissue and tumours with 99% success compared with 77% pre-operative diagnosis for the same patients which was based on all the available clinical information including CT, MRI and angiography MRI [Preul *et al.*, 1994] [Hagberg *et al.*, 1995]. MRS has also been shown to be useful in the diagnosis and grading of prostate cancer [Heerschap *et al.*, 1996]. For a review of studies of human tumours by MRS see [Negendank, 1992]. Investigators studying response to therapy (for example radiotherapy and chemotherapy) in animals have found that the ^{31}P spectrum changes in response to therapy often before there is a noticeable decrease in tumour size [Leach, 1992] [Weiner, 1994].

Another example of a potential application of MRS is in the treatment of epilepsy. MRS is already used in some centres to help identify the focus of seizures before brain surgery. Current EEG and other scanning methods do not provide accurate localisation information in the majority of cases, but it has been shown that ^{31}P and ^{1}H provide additional information that may avoid the use of invasive depth electrodes [Weiner, 1994].

MRS may also be useful for the evaluation of metabolic myopathies. ^{31}P MRS is already used at a number of centres to determine the presence of metabolic myopathies from elevated inorganic phosphate peaks in resting muscle [Weiner, 1994].

Another possible useful application of ^{1}H MRS is in the prediction of outcome within the first few months of life for new-born babies with hypoxic-ischaemic encephalopathy [Peden *et al.*, 1993].

Examples of other diseases where *in vivo* MRS has potential clinical application are AIDS, Alzheimer's disease, stroke and MS. Several comprehensive reviews have been written on the potential application of MRS in clinical medicine, for example [Ross and Michaelis, 1994] [Bot-

tomley, 1989] [Weiner, 1988] [Vine, 1990].

2.6 Spectral Analysis

2.6.1 Problems Associated with *in vivo* Spectral Analysis

One of the great advantages of MRS for medical applications is that it allows us to obtain information about the metabolic composition of living tissues *in situ*. However, the fact that these signals are obtained *in situ* presents considerable difficulties, both with acquiring the signals and extracting the relevant information. The fact that it is impossible to control all the conditions of an examination carried out *in vivo* means that the signal may contain unwanted artefacts, for example those due to the movement of the patient, which effectively changes the sample which contributes to the MR signal. Another problem is that while it may be possible to focus on a specific region it is often not possible to focus on a specific tissue. Because the size of smallest region that can at present be examined effectively by human *in vivo* MRS is approximately 2 cubic centimetres it is likely that signal from any region will include other signals in addition to those from the required tissue, for example signals from neurons and blood cells [Bock, 1994].

Some of the problems with spectral analysis affect experiments carried out both *in vivo* and *in vitro*. One such problem occurs when the number of metabolites that can be observed is large. This leads to a crowded spectrum with many possibly overlapping peaks. If these peaks are too close together, it may be difficult both to identify and to subsequently quantify them. This problem may be particularly severe for ^{1}H spectra where the ppm range is relatively narrow (8 ppm).

Another problem is due to the fact that MRS is "not exceptionally sensitive" [Spisni, 1992]. The sensitivity, which can be expressed in terms of the signal-to-noise ratio of the spectrum, is dependent on several factors. These include the strength of the applied field B_0, the design and performance of the NMR instruments and the time taken to accumulate the data. One of the main factors that accounts for the low sensitivity of NMR is that the interaction between the nuclei and the magnetic field is weak, that is the amount of energy absorbed is low. This means that the amount of energy released is also low leading to a weak signal. Different nuclei have different sensitivities. Also the abundance of a certain nuclear isotope may be low, e.g. ^{13}C which has a natural abundance of only 1.1%.

Because of the limitations imposed on acquiring a signal from a living subject, the signal-to-noise ratio is generally lower for data acquired *in vivo*. The signal-to-noise ratio depends on the strength of the magnetic field and the high magnetic fields used for *in vitro* studies are not permissible for living subjects. Also, the low signal-to-noise ratio cannot be improved using averaging techniques because of the short experimental times required for patient comfort. While it is possible to enhance the signal- to-noise ratio using apodisation, this will result in line broadening and so more peak overlap. Lines tend to be broader in any case due to the magnetic field inhomogeneity caused by the heterogeneous nature of the sample which may also cause distortions in the lineshapes.

Baseline distortion is another problem which may affect quantification of *in vivo* data. One factor which can alter the shape of the baseline is the presence of metabolites with large peaks which have broad 'humps' which spread either side of these peaks. The effect of these humps is to 'push up' the contributions of other metabolites in the spectrum. This is a problem which particularly affects kidney, liver and tumour ^{31}P spectra, where a broad hump of signals from immobile phosphates underlies the spectrum [Andrew *et al.*, 1990]. Another cause of baseline distortions are a side-effect of chemical shift imaging when a few datapoints at the beginning of

the FID are lost due to the delay between the pulse and the commencement of acquisition. This has the effect of dephasing the spectrum, causing 'wiggles' in the baseline. In theory it should be possible to use first order phase corrections, to remove these wiggles but in practice this is often not possible [Wang, 1992] [Saeed and Menon, 1993].

Another factor which will affect spectral analysis methods is that while in principle, MRS spectra should have Lorentzian peaks, the peaks observed from spectra obtained *in vivo* are often not of this ideal shape (equation 2.3). This may be due to magnetic field inhomogeneity, magnetic susceptibility and other problems. Even when the peaks are Lorentzian, their wings stretch to infinity and even taking into account a total wingspan of $\pm$ 3.2 times the width at the half height, only 90% of the peak area will be accounted for.

In addition to the problems outlined above, it is also necessary to take operator dependent factors into account. For example, the method of phase correction of the spectrum directly influences the shapes and heights of the peaks, and will affect any subsequent analysis of the spectrum. It is thus important to make sure that the spectrum is correctly phased. Since interactive methods of phasing vary according to the judgement of the operator, this is best achieved using an automated procedure.

2.6.2 Methods of Spectral Analysis

Semi-automated Methods

Spectral analysis is usually carried out by first identifying and then measuring the peaks representing metabolites of interest in individual data. Since absolute quantification is often not possible for data acquired *in vivo*, the measurements are usually of relative concentrations. Metabolite concentrations are obtained from frequency domain data by identifying and selecting the peaks and then measuring these peaks. While software is available on most spectrometers for spectral analysis, the process is normally an interactive one. Most software for spectral analysis requires interaction by the user at least three times; firstly for phase adjustment, then for selecting the peaks that the analyst thinks are important and thirdly for flattening the baseline to facilitate integration of the areas under the peaks.

There are a number of methods available for measurement of peak areas, including printing the spectrum on paper, cutting out the peaks and weighing them. If the peaks are all the same width and shape, relative concentrations may be deduced from peak heights. Since this is not usually the case, curve fitting methods are normally used. These involve the fitting of analytical curves (normally Lorentzian or Gaussian) to the peaks by least squares methods.

Automated Methods

Several alternative approaches have been taken to support the full automation of MR spectral analysis. One common approach is to find ways of quantifying the metabolites of interest directly from the FID signal. This is generally done by fitting a specific model function to the original signal. These methods may use prior knowledge of the data, for example expected resonance frequencies, or interpolation parameters. The output will be numeric values for the parameters of interest. See [Wang, 1992] [de Beer and van Ormondt, 1992] [van Dijk *et al.*, 1992] [Diop *et al.*, 1992] [Joliot *et al.*, 1991] for a full description of these methods. Another approach is to fit the entire Fourier transformed spectrum (as opposed to individual peaks) to a model. For example Provencher [Provencher, 1993] has developed a method, the LC method, by which an *in vivo* spectrum is analysed as a combination of model spectra of metabolite solutions *in vitro*.

Several other interesting methods have been proposed for spectral quantification, which have yet to prove their worth. Example are the use of wavelets for quantifying overlapping resonances in the time domain data [Serrai *et al.*, 1995], and PCA for quantifying individual peaks [Stoyanova *et al.*, 1995].

The Pattern Recognition Approach

The disadvantages of spectral analysis based on interactive peak quantification are obvious. Firstly such measurements may not be accurate, owing to problems such as baseline distortions and overlapping peaks which make quantification very difficult. In addition they require subjective judgements by the operator, and the results from different operators may be very different. In addition they are very time-consuming. Fully automated methods are clearly desirable and the newer methods such as time-domain fitting can overcome some of the problems. However most of these still require that the metabolites of interest are specified in advance, and thus only information from a specified part of the spectrum is used.

The methods that have so far been discussed for spectral analysis have the specific purpose of quantifying metabolites. An alternative approach is to treat spectral quantification as a process of extracting measurements that are directly related to differences between spectra, rather than differences between metabolites. This approach uses the whole spectrum without assigning prior importance to any peaks or ppm regions of the spectra and considers groups of spectra simultaneously rather than individually. This approach is complementary rather than an alternative to the other approaches as it has a different aim. It is sometimes referred to as 'the pattern recognition approach', as the aim is to find patterns in the data based on the spectral datapoints, rather than measured metabolite ratios. In fact pattern recognition methods can be applied to any measurements including peak intensities, but feature extraction, that is the identification of the most suitable measurements for classification is a major part of the pattern recognition process.

Since the aim of this thesis is to develop automated methods that use information from the whole spectrum which use no preconceptions about position or importance of metabolites in the spectrum this approach is particularly suited to the purpose of this thesis. The methods that were used, and the results of applying these to the data that was studied are described in detail in the following chapters.

Chapter 3

Analysis of *in vivo* Magnetic Resonance Spectra Using Pattern Recognition

---

3.1 Introduction

> 'An important part of scientific activity consists in gathering data which are mostly the result of measurements. In fact, modern analytical chemical and physical measuring methods provide an ever increasing amount of information. In medicine, clinical examination and complementary investigation, e.g. biochemical analyses, result in a large amount of data which allow the investigator to draw as complete as possible a picture of the physiological (normal or abnormal) state of the patient. Whereas the assembling and storing of data has steadily increased since the availability of modern computer data-acquisition methods, proper **target-interpretation** of these data has received poor attention so far, resulting in a rather poor utilisation of the available information [Coomans and Broeckaert, 1986].

Although the passage quoted above was published in 1986, it could have been written about clinical MRS in 1996! Recent advances in technology for acquiring and processing MR signals *in vivo* have made it possible to obtain important information about the physiological composition of living tissue. However, while much effort is being applied to the development of improved techniques for the acquisition and processing of MRS data and subsequent measurements of metabolite concentrations, relatively little effort has been applied to the development of methods for the interpretation of these measurements once they have been acquired. Because of this, the potential information available from the spectrum is often not fully utilised.

MRS provides a means of non-invasively observing the biochemistry of living tissue *in situ*. The two main advantages of MRS for clinical medicine compared with most other methods of chemical analysis are:

- Because the signals carry information on a large number of compounds, MRS provides a means of investigating the metabolic composition and processes of living tissues [Gadian, 1995].
- Because NMR is non-invasive, MRS signals can be acquired from living tissue without changing the nature of the tissue and without harming the patient.

These advantages make MRS an extremely useful tool both for clinical medicine and for studying disease. However associated with these advantages are problems that can make it difficult to obtain and extract relevant information from MRS data. It is not possible to control all the conditions of an examination carried out *in vivo*. The tissue being examined will not be homogeneous and the signal may contain unwanted artefacts. These factors together with a low signal-to-noise ratio may make identification and measurement of the metabolites that are present in the tissue very difficult. The fact that MRS signals carry information on a large number of metabolites presents the non-trivial problem of how to extract and utilise this information, i.e. how do we use the data to build up a picture of the metabolic composition of the tissue?

In this chapter, I discuss some of the ways that pattern recognition analysis might be used to help overcome some of these problems, by providing methods for target interpretation of MRS which make the best use of the available information and which answer the type of questions that will need answering if MRS is to play a useful role in clinical medicine and also by providing alternative methods for extracting features from the spectra.

3.2 The Advantages of Pattern Recognition Analysis for MRS

The term pattern recognition encompasses a wide range of techniques for analysing and interpreting complex data [Duda and Hart, 1973] [Fukunaga, 1990] [McLachlan, 1992] [Coomans and Broeckaert, 1986]. The aim of pattern recognition is to find patterns in data which can be used to discriminate between subgroups of the data and to identify important distinguishing factors. Many computer-based pattern recognition applications are directed at finding ways of automating processes that humans do naturally, such as understanding language, or interpreting visual scenes. However pattern recognition techniques also provide a means of extracting relevant information from complex data that humans find difficult to interpret. In this case the emphasis is on helping the human analysts than rather than on replacing them. Pattern recognition techniques are widely used and are very applicable to analytical chemistry, where large amounts of data are often involved. They are particularly useful for spectral application where the number of measurements obtainable from a single sample may be very large. This includes the analysis and interpretation of non-medical MRS data, where reported applications of pattern recognition analysis include helping to elucidate the structure of chemical molecules, and classifying samples according to their MR spectra, for example [Kowalski and Reilly, 1971] [Kormos and Waugh, 1983].

There are a number of reasons why the pattern recognition approach is very appropriate for the analysis of medical MRS data.

- PR is a multivariate approach which uses not only the information contained in each single variable or measurement, but also information drawn from the relations between the variables [Coomans and Broeckaert, 1986]. MRS data often contains information on many metabolites and the multivariate approach provides a means of exploring the relationships between these metabolites.

- Pattern recognition provides methods for discriminating between samples of different classes and for assigning a sample to a particular class. These methods will often include estimations of the probability of a certain sample belonging to a certain class.

- Pattern recognition techniques facilitate the presentation of multi-dimensional data in a form that can be easily viewed, e.g. simplified two-dimensional displays. These displays can be used to investigate unknown groupings or to find the best patterns to separate the groups.

Alternatively they may be used to show how typical or atypical a spectrum may be compared with others in the group.

- Pattern recognition is a computer-based approach, which can deal with a large amount of data and is well suited to the analysis of large and complex data files. Automated methods are desirable, not only because they are far less time consuming, but also because they offer an objective and unbiased method of analysis.
- Pattern recognition, because it is concerned with finding the best patterns to discriminate between classes of data, can provide alternative methods for quantifying spectra using features other than explicit peak measurements. This can be very useful for spectra acquired *in vivo*, for which traditional methods of quantification may be problematic.

3.3 Pattern Recognition: Basic Approaches and Definition of Related Terms

Pattern recognition is a very large subject which draws together methods from various related disciplines. This section gives a brief summary of the different approaches of pattern recognition, and definitions of related terminology.

There are three different approaches to pattern recognition [Schalkoff, 1992]. Historically, the two main approaches are the statistical (or decision theoretic) and the syntactic (or structural) approaches. More recently there has been a wide interest in the third approach of using neural networks for the 'black box' implementation of pattern recognition algorithms. These three approaches are not mutually exclusive, for example some neural classifiers, e.g. probabilistic neural networks, are designed to implement more traditional statistical algorithms [Masters, 1995]. All three approaches have the same aims, that is to describe the important features in the data and to provide methods for discriminating between different types or classes of data.

3.3.1 The Statistical Approach

The statistical pattern recognition approach is based on the statistical study of measurements made on the data to be classified. A set of characteristic measurements, denoted features, are extracted from the input data. These features, which may be a subset or combination of the original measurements, are expressed as a vector $X = (x_1, \ldots, x_n)$ in the n dimensional feature space R^n. The problem of assigning a feature vector to a particular class is tackled by estimating density functions in the n dimensional space, and dividing the space into regions of categories or classes [Schalkoff, 1992] [Miclet, 1986]. Ideally different class populations will occupy different regions in the feature space allowing classification methods to allocate test observations based on their location in the space [Aeberhard *et al.*, 1994].

3.3.2 The Structural Approach

The structural pattern recognition approach provides methods for describing the data in a manner related to its structure. Rather than representing the features as 'meaningless' numbers, as does statistical PR, it attempts to provide structural descriptions in which intrinsic characteristics of the features (for example the shape) are taken into account. Rules are formulated to compare these structural descriptions and to characterise similar structures by summarising their structural properties [Miclet, 1986]. Structural pattern recognition is often called syntactic pattern recognition since the structures of the patterns are often related to the syntax of a formally defined language. Syntactic PR draws on the vast body of knowledge that has been acquired in the study of both

natural language and programming languages. As in language theory, patterns are described as sentences which are analysed by parsing. Typically syntactic PR approaches formulate hierarchical descriptions of complex patterns built up from simpler sub-patterns [Schalkoff, 1992].

3.3.3 The Neural Approach

The neural pattern recognition approach uses artificial neural systems termed neural networks for classifying data. Neural networks are computer-based systems which model the way that biological neural systems manipulate information. Neural networks have the ability to handle large amounts of data, and to form rules and discover patterns within them. The basic design is very simple: a neural network consists of a number of 'nodes' which emulate the neurons. Each node receives a number of inputs. The feature vector provides the input to the first layer of units. For subsequent layers, the inputs come from the outputs of units in the preceding layer, which are modified by weighted connections. The output of a unit is determined by the sum of its inputs and a threshold function. The weights are iteratively adjusted using a learning algorithm to optimise the output of the network according to some cost function. Typically a network will consist of two or three internal layers (hidden layers) [Bishop, 1995].

3.3.4 Multivariate Analysis

The methods used for statistical pattern recognition form a subset of methods used for multivariate data analysis which is the application of statistical techniques to multivariate data. While some multivariate techniques are extensions of univariate statistical techniques, for example multiple regression methods, or multiple analysis of variance, the majority of multivariate techniques have been developed to deal with the special problems of dealing with multidimensional data, and thus may be very appropriate for pattern recognition problems. Many of these methods are devoted to combining or reducing the variables while keeping the basic structure of the data intact (reduction of dimensionality) [Krzanowski, 1988] [Everitt and Dunn, 1991].

3.3.5 Chemometrics

'Chemometrics', a term coined in 1972, can be defined as 'the chemical discipline that uses mathematical, statistical and other methods employing formal logic to design or select optimal measurement procedures and experiments, and to provide maximum relevant information by analysing chemical data' [Massart *et al.*, 1988]. Chemometrics encompasses a large number of methods for pattern recognition analysis and signal processing of chemical data, many of which are very appropriate for spectral data, including MRS data. All of the methods that have been adopted in this thesis can be described as chemometric techniques. Massart et al. give a thorough description of virtually all of the currently known techniques appropriate for pattern recognition analysis of spectra in their excellent textbook on the subject [Massart *et al.*, 1988].

3.4 Approach Used in this Research

The decision to use statistical pattern recognition analysis was taken at a fairly early stage of this research. Both the statistical or neural approaches are appropriate for classifying numerical data such as spectra. The structural approach could prove useful in identifying differences between classes of spectra that are based on the shapes of the peaks, and this approach was considered at an early stage of this work when a method of describing shape was required. However this was not

pursued when it became apparent that the wavelet transform (see section 3.5.3) provided a quick and successful method for condensing spectral shape information.

The two main reasons for choosing the statistical rather than the neural approach were:

- Statistical pattern recognition methods are based on firm theoretical foundations. Many of the techniques have been developed and tested over a long period of time. By contrast artificial neural networks have only recently been applied to data analysis problems. It seems that most classification problems can be handled equally well by statistical techniques as by neural networks, and therefore there is no advantage in using them.
- In general neural network classifiers take much longer to train than do classical methods of discriminant analysis. This was an important consideration given that the main task of this research was to select the best features for discrimination from a large number of potential features.

Both the statistical and neural approaches have been used for previous PR studies of MRS data. Use of neural networks for classifying medical MRS data is described, for example by [Howells *et al.*, 1992a], [Anthony *et al.*, 1994] and [El-Deredy and Branston, 1994] and for identification of important spectral features by [Friesen *et al.*, 1995] and [El-Deredy *et al.*, 1995].

The following section discusses methods of statistical pattern recognition that may be suitable for describing and classifying medical MRS data. This section includes reference to examples of reported studies applied to medical MRS data. While most of the methods that are described are standard statistical PR techniques that may be applied to all types of data, a few are particularly useful for chemical data such as spectra. While details of the former, for example discriminant analysis and PCA, will be found in any text book on statistical PR, details of the latter, for example variance weighting, or SELECT, are generally only to be found in the literature on chemical PR or chemometrics.

3.5 Statistical PR: Appropriate Methods for Analysis of MRS Data

The fundamental idea of pattern recognition is that when a sample is characterised by a number of measurements, these measurements form a pattern. This pattern can be used to answer three questions [Massart *et al.*, 1988]:

- how to classify an object in one of two or more known classes on the basis of its pattern
- how to detect groups of samples with similar patterns and
- how to display multidimensional data in a lower and preferably two-dimensional space without significant loss of information.

Methods for answering the first question are generally grouped together under the general heading of discriminant analysis or ‘supervised pattern recognition’. These classification methods rely on having some data of known class available to ‘train’ the system, that is to find the best ways of partitioning the feature space into class regions, and to develop rules for assigning new cases to a specific class. In order to test the system it is usual to set aside some of the data whose class is known during the development of the system and then to use this ‘test set’ to evaluate the performance of the classification rules. There are a number of methods for evaluating the

discriminant rules, depending on how the test set is chosen. If there are a large number of samples it is usual to divide them into two groups and use one for training and the other for testing. If, however there are only a few samples of known class it may be necessary to use what is known as the 'leave one out' method for assessing the success of the discriminant rules. This method entails using all the cases except one as the training set, and then using the excluded case as the test set, repeating this process until the whole set has been tested.

The problem of how to detect groups of samples with similar patterns is tackled using cluster analysis or 'unsupervised pattern recognition'. In this case the class structure of the data will be unknown, and there will therefore be no data of known class available to train the system. Instead, the system will try to detect natural groupings or clusters in the data.

For the third question the main problem is how to reduce the dimensionality while keeping the basic structure of the data intact. If the number of variables representing the data can be reduced to two or three, display techniques such as two and three dimensional scatter plots can be used to allow the observer to discern the class structure and grouping within the data. Such displays can be used to assist the analyst for classifying new data and also for finding clusters within the data. Indeed, sophisticated computer classification and clustering methods may be unnecessary for data consisting of two or three variables since groupings can normally be discerned by eye.

Reduction of dimensionality is a key problem in pattern recognition, particularly in applications where the number of variables is high compared with the number of samples, as is often the case with medical MRS data. Most classification methods depend on a certain ratio of samples to variables; normally the number of variables should be no more than one third the number of samples [Kowalski and Wold, 1982]. One of the recurring problems encountered in applying statistical techniques to pattern recognition problems is due to 'the curse of dimensionality' [Bellman, 1957]. Procedures that are analytically or computationally manageable in low-dimensional spaces can become completely impractical in a space of 50 or 100 dimensions [Duda and Hart, 1973]. In this research, reduction of dimensionality, that is how to extract a small number of variables from the vector of spectral datapoints, was a major concern.

This section is structured as follows: the first part on discriminant analysis discusses methods for the classification of data, the second part discusses clustering methods for detecting groups of samples with similar patterns, and the third discusses methods for addressing the key question of reduction of dimensionality of the data.

3.5.1 Discriminant Analysis

The major role of discriminant analysis is to define criteria (classification rules or decision rules) for classifying test objects into one of the training classes. If the objects are represented by vectors in n-dimensional space, i.e. $(X = (x_1, \ldots, x_n))$, each object can be thought of a point in this n-dimensional space. Geometrically the formulation of a classification rule corresponds to an explicit or implicit construction of a boundary surface between the training classes so that the classes become as well separated as possible. In this way the pattern space is divided into as many regions as there are training classes in the training set [Coomans and Broeckaert, 1986].

The construction of such a boundary can be found by optimisation, by starting with a boundary chosen at random and iteratively shifting it until the boundary which best separates the classes is found. This procedure is often referred to as the 'linear learning machine'. In this technique each object is classified into one class with no measure of doubt. This is unsatisfactory for many applications and most statistical discriminant methods use probability theory to estimate the possibility of an object belonging to a certain class. These are called probabilistic or Bayesian methods

of discriminant analysis. Using probabilistic methods boundaries are constructed by estimating density functions in the n-dimensional space, and then deriving a rule for allocating each object to a certain class. If the true probability density functions of all the classes are known, the optimal decision rule for classifying an unknown object X is to allocate it to class i if

$$P(X|g_i)P(g_i) \; > \; P(X|g_j)P(g_j) \; for \; all \; j \neq i \tag{3.1}$$

where $P(g_i)$ is the probability of an object belonging to group i and $P(X|g_i)$ is the probability of getting a set of measurements X given that that object belongs to group i. This rule is optimal if the true probability density functions of all the classes are known. However in practice the densities $P(X|g_i)$ will be unknown and must be estimated from the training data.

The above rule is derived using Bayes rule (3.2) which provides a method for estimating the probability that a sample with feature vector X belongs to group i.e. $P(g_i|X)$. This is the probability that we really need, but which is much more difficult to to find by standard methods of estimation than $P(X|g_i)$.

$$P(g_i|X) = \frac{P(X|g_i)P(g_i)}{\sum_{i=1}^{g} P(X|g_i)P(g_i)} \tag{3.2}$$

where g is the number of groups [James, 1985] [Fukunaga, 1990].

Parametric methods of density estimation model the classes using assumptions about the underlying probability densities of the data set (usually assuming a multivariate normal distribution). The training samples are used to estimate the parameters in these models. Examples of parametric classification methods are quadratic and linear discriminant analysis. Nonparametric methods make no such assumptions. Instead the test objects are classified on the basis of the training samples in the neighbourhood of the object in the feature space. Examples of nonparametric classifiers are the K- nearest neighbour and potential methods [Aeberhard *et al.*, 1994].

The advantage of parametric methods is that the required calculations are simplified and need relatively little processing time. Nonparametric methods rely on densely populated feature spaces for reliable classification and require a large number of samples. Apart from the fact that it is extremely difficult to obtain an accurate density estimation in high-dimensional feature space, nonparametric methods are normally much more time-consuming than parametric methods of density estimation, particularly when there are a large number of variables. However, it is important to be aware that the use of a simple parametric model, such as the multivariate model may give misleading results if the assumption about the distributions are invalid.

Parametric Discriminant Analysis

The most widely used statistical classification method is linear discriminant analysis (LDA). In this method the decision as to whether an object should be allocated to a particular class is made on the basis of its discriminant score or scores.

A discriminant score, which is a weighted linear combination of the original variables is calculated using a discriminant function

$$D(X) = w_0 + w_1x_1 + w_2x_2 \ldots + w_nx_n \tag{3.3}$$

Where the w's are constants (or weights) which, for the two class case, are found by maximizing

$$\frac{\sum_{j=1}^{n} w_j(\overline{x}_{kj} - \overline{x}_{lj})^2}{\sum_{j=1}^{n}\sum_{j'=1}^{n} w_j w_{j'} c_{jj'}} \tag{3.4}$$

where $\overline{x}_{kj}$ is the mean of variable x_j for class k and $c_{jj'}$ is an element of the pooled or average variance-covariance matrix [Massart *et al.*, 1988]. This has the effect of minimizing the within-class variance and maximizing the between-class variance, thus minimizing differences within the groups and maximizing differences between groups. The discriminant function can be used as a classification rule for allocating an individual to a particular group. Each individual is assigned a discriminant 'score' which is the weighted combination of its values of the discriminating variables given by equation 3.4. The decision as to whether the particular individual comes from one group or another is based on measuring the distance between its particular score and the centroids of the two different groups, and comparing the probabilities of its membership of each class. If the prior probabilities, $P(g_i)$, i.e. the probability of an object belonging to group i, are known these can be incorporated into the discriminant rule. For example these probabilities might be approximated from knowledge of the relative sizes of the classes. When little is known about the relative sizes it is usual to assume they are equal.

The linear discriminant rule is often referred to as Fisher's discriminant rule [Fisher, 1936]. This rule is optimal if the populations come from a multivariate normal distribution and have equal covariance matrices. However Fisher derived the method without directly assuming that the probability density functions were normal and it therefore can be expected to perform reasonably well even when the assumption of normality is not wholly justified as is often the case [Everitt and Dunn, 1991]. When there are more than two classes it is possible to determine several combinations of the original variables (called canonical variates) for separating the groups. The first discriminant function, as in the two group case, has the largest ratio of between class variance to within class variance; the second, which is uncorrelated with the first will have the next largest ratio etc. In general if there are k groups, $k-1$ functions may be computed [Norusis, 1994]. Since the assumptions of normality and equal covariance only influence the discriminant boundaries but not the discriminant function, these functions can be useful for displaying the data.

Quadratic discriminant analysis uses the same principles to derive the discriminant rule as LDA. However no assumption is made as to the equality of the covariance matrices of the different groups. Since this assumption has the effect of simplifying the decision rule and reducing the number of parameters to be estimated, QDA is a more complicated procedure than LDA. When the covariance matrices are taken to be unequal, the discriminant functions include a quadratic term, meaning that the subsequent decision boundaries are quadratic, rather than linear as for LDA. Quadratic and linear discriminant analysis are well-known standard techniques, which are described in most books on multivariate analysis or statistical pattern recognition, e.g. [Fukunaga, 1990], [McLachlan, 1992], [Everitt and Dunn, 1991].

Linear discriminant analysis is a popular technique, which is applicable to many types of data and which is provided by all the standard multivariate software packages. It is popular because it is a computationally simple technique which is relatively robust to departures from the assumptions of normality and equal covariance matrices for the different classes. LDA has been extensively used and discussed in the literature on statistical PR. A comprehensive discussion and comparison of LDA with other classical discriminant techniques for classifying a data set of patients with head injuries, is given in the seminal paper by [Titterington *et al.*, 1981] and the discussion that follows the paper. This paper shows along with others such as [Aeberhard *et al.*, 1994] and [Sjostrom and Kowalski, 1979] that in general, LDA performs well in comparison with other techniques. This has also been shown to be true for MRS data. [Nikulin *et al.*, 1995] state that in their extensive experience of classifying spectra obtained from various biopsies, that if the pre-processing is carried out correctly, "accurate classification then follows even with simple classifiers such as LDA".

In analytical chemistry, linear discriminant analysis has been used both to classify unknown samples and to identify important characteristic chemical features of groups of data. Its use for the

chemical analysis of MR spectra was reported as early as 1971 when Kowalski and Reilly used LDA to develop a classification rule to distinguish between ethyl, n-propyl and iso-propyl groups from a training set of ^{1}H spectra [Kowalski and Reilly, 1971]. In this study the whole spectrum rather than selected frequencies was used and the spectra were pre-processed using autocorrelation methods before classification. LDA has been used by a number of groups applying PR techniques to medical MRS data, obtained both *in vitro* and *in vivo*. [Preul *et al.*, 1994] and [Hagberg *et al.*, 1995] used LDA to classify glial brain tumours on the basis of metabolite measurements i.e. peak measurements from ^{1}H spectra obtained *in vivo*. In both studies the tumours were divided into three grades on the basis of biopsy data and good separation was obtained between the three groups. A group from the Institute of Biodiagnostics in Winnipeg have used LDA in a number of studies using MRS data of various biopsies. They have shown that the technique can be used to successfully classify ^{1}H spectra of various diseases including thyroid neoplasms [Somorjai *et al.*, 1995b], cervical dysplasia [Nikulin *et al.*, 1995] [Friesen *et al.*, 1995], human brain neoplasms [Nikulin *et al.*, 1995], and colorectal cancers [Somorjai *et al.*, 1995a].

Nonparametric Discriminant Analysis

The nearest neighbour method The nearest neighbour method is one of the simplest nonparametric method of classification. In this method the decision for assigning an object to a particular class is made by comparing its measurement vector with each of the vectors of the training set. The distance (generally the Euclidean distance) between the object's vector and each vector in the training set is computed and the lowest of these is selected. In a more sophisticated version of this, called the k-nearest neighbour method, the k nearest samples are selected and the object is allocated to the class to which the majority of the k samples belong. Although this method is mathematically simple, the computational cost of calculating the distance of every vector in the training set can be very large [Duda and Hart, 1973].

Kernel Methods Kernel methods (also called potential methods) of density estimation are nonparametric decision methods. They differ from the parametric methods of density estimation in that the conditional probability densities are not assumed to come from a known parametric family [Silverman, 1986]. Instead the shape of the probability distribution function of a given class is estimated on the basis of measurements of the training objects and by means of direct density estimation [Coomans and Broeckaert, 1986]. The density function is estimated using kernel function K which satisfies the condition $\int K(x)dx = 1$. Usually (but not always) K will be a symmetric probability density function such as the normal density. The kernel estimator with density K is defined by

$$\tilde{f}(x) = \frac{1}{mh} \sum_{i=1}^{m} K(\frac{x - X_i}{h}) \tag{3.5}$$

where h is the *window width*, m is the number of samples and $X_1 \ldots X_m$ are the values for each sample [Silverman, 1986]. K acts as the shape parameter and h is a smoothing parameter. For most applications K is fixed while h is specified as a function of the data [McLachlan, 1992]. The advantage of these methods, as with any nonparametric method are that no assumptions are made about the probability density functions. They should therefore perform better if the data does not come from the distribution that has been assumed by the parametric methods. The disadvantage is that a large number of calculations may be needed to estimate the density functions for high-dimensional data. Also the choice of the smoothing parameter h can strongly affect the performance of the classifier.

To my knowledge, kernel methods of density estimation have not been used for classifying medical MRS data, probably because they need a densely populated feature space to estimate the probability distributions, thus requiring a large number of samples. The use of kernel methods of

density estimation in chemistry has not been widely reported. [Coomans and Broeckaert, 1986] give a very comprehensive review of medical and chemical applications of these methods. The nonparametric methods in general need a much larger sample-to-variable ratio and are not practical for most application of medical spectral data analysis where this ratio is generally low. However, if spectroscopy becomes more widely used as a diagnostic tool and more data is acquired, it may be found that these methods are useful.

Other Supervised Learning Techniques

There are a number of other supervised learning techniques suitable for classifying chemical data, for example classification and regression trees (CART) [Breiman *et al.*, 1984]. In this method the data are split into two parts based on the value of variable x_1, which minimises interclass variance and maximises between class variance. After splitting on one variable the separate parts are then split again. Variable x_2 may be used to split one part, and perhaps x_3 or x_2 or even x_1 again may be used to split the other part. Another class of techniques particularly suitable for classifying chemical data are the so called 'modelling' techniques such as SIMCA in which a model is developed for each class, independently from other classes in the training set using principal components. For a description of these and other methods see [Massart *et al.*, 1988].

3.5.2 Cluster Analysis

Cluster analysis is the generic term for a large group of unsupervised learning techniques which attempt to identify natural groupings in a data set. The simplest approach to discovering distinct groups or clusters is by examination of scatterplots. If there are only a few variables this can be done by plotting two or three of these at time to see whether there are any obvious groupings. Otherwise the first two or three principal components or the results from multidimensional scaling may be plotted (see section 3.5.3).

The most widely used strategy for computer-based cluster analysis is hierarchical clustering. Hierarchical clustering can be agglomerative or divisive. Agglomerative clustering proceeds sequentially from the stage in which each object is considered to be a single member cluster to the final stage in which there is a single group containing all n objects and divisive clustering starts with the whole group and successively splits this group into a number of clusters [Schalkoff, 1992]. The results can be displayed as a dendrogram. There are a large variety of clustering methods which differ depending on which criteria are used for defining inter-group distance or similarity. For details of these methods see for example [Krzanowski, 1988] or [Everitt and Dunn, 1991].

Cluster analysis can prove very useful in exploratory data analysis. However it can pose problems, one being that different methods may provide completely different clustering in the same set of data. Also, while it may be easy to find clusters in data it may not easy to give meaning to the groupings. It is important to be aware that clustering techniques will generate a set of clusters even when applied to random, unclustered data. Other methods for discrimination are normally preferable when the groupings are already known [Krzanowski, 1988].

Cluster analysis works well when the groups are very well separated and may therefore be a useful method for initially checking the data. For example 'rogue' data which have been entered into the analysis by mistake, or alternatively pre-processed incorrectly may be detected using this method. For example, it proved very useful in this research for detecting two samples which had erroneously been labelled with the wrong acquisition time.

Gartland et al [Gartland *et al.*, 1991] applied cluster analysis to ^{1}H urinanalysis data from a

variety of induced toxic states in rats. Hierarchical cluster analysis was used as a preliminary method of analysing the intensities of 16 metabolites obtained from these spectra. The cluster dendrogram showed that some of the different toxins formed a discrete cluster.

Howells and co-workers have used cluster analysis to categorise ^{1}H spectra obtained from perchloric acid extracts of normal and tumorous tissue in rats [Howells *et al.*, 1992a] [Howells *et al.*, 1992b]. The spectral datapoints were processed using a specially developed digitisation technique (see section 3.5.3) and principal component analysis was used to reduce the dimensionality of the data further. The dendrogram showed a partial separation of the samples into groups representing the different tissue types. In another study by the same group similar techniques were employed to cluster a set of *in vivo* ^{31}P animal data. The resulting dendrogram showed that most of the groups formed distinct clusters [Howells *et al.*, 1993a]. Cluster analysis was also used by [Hagberg *et al.*, 1995], but LDA gave better separation between the groups than the clustering algorithms.

3.5.3 Reduction of Dimensionality

Many applications of pattern recognition in chemistry (including spectroscopic applications) involve a large number of variables and often the first problem to be dealt with is how to reduce these in some way. The main reasons for this are as follows:

- to facilitate visualisation of the data, and to allow the analyst to discern class structure and groupings within the data,
- to reduce the number of variables for classification, either in order to reduce the ratio of variables to samples, or to reduce the computational complexity of estimating the density functions.
- to find salient features in the data, that is to find the best features or combinations of features to represent (and possibly explain) differences between different classes of data.

The process of reducing the number of variables, while keeping the basic structure of the data intact, is often called feature extraction. Dimensionality can be reduced by selecting appropriate subsets of the available variables or by combining the variables in some way. Methods for reduction of dimensionality fall into two categories. In the first category are methods which aim to describe the data more succinctly, that is to express the data as concisely as possible with minimum loss of information. These methods do not rely on prior knowledge of class membership of the subjects but attempt to sift out irrelevant information by transforming the original into a new set of variables with the hope that a proportion of the new variables can be discarded with little loss of information. Examples of such methods are PCA or the wavelet transform. The second category of methods attempt to reduce the number of variables by selecting the best set of features for discrimination. In this case knowledge of the class membership of the data will be used. Often a combination of the two types of methods may be used to find the best discriminating features.

Reducing the dimensionality for Data Visualisation

> "A complete analysis of multidimensional data requires the application of an array of statistical tools – parametric, nonparametric and graphical. Parametric analysis is the most powerful, nonparametric is the most flexible and graphical analysis provides the vehicle for discovering the unexpected" [Scott, 1992].

Pattern recognition techniques provide ways of extracting relevant information from complex data that humans find difficult to interpret. The previous two sections discussed how this can be

done using computer-based methods for discriminating between different classes or for finding groupings and clusters in data. These methods are needed if the data has many variables, since it is very difficult for the human analyst to visualise and interpret high-dimensional data. However, if the dimensionality can be reduced to two or three variables it may be possible to discern natural groupings in the data by eye, after plotting these values on a scatterplot. We can also see whether these groups are linearly separable, that is whether a straight line or, in the case of three dimensional data, a surface can be 'drawn' between them, or alternatively if any other (non-linear) decision boundaries might be constructed

Apart from being very useful in identifying meaningful groupings, reduction of dimensionality for data display allows preliminary exploration of the data before further analysis. Scatterplots are useful for finding unknown groupings of the data and also for detecting outliers, which if not eliminated might affect the subsequent analysis. They are also useful for showing the results from a classification or cluster analysis. For example by displaying the scores obtained from discriminant analysis on a scatterplot the typicality of each individual sample compared with the others can be demonstrated and may aid the user in a decision.

Reduction of variables for display can be achieved either by selecting two or three of the original variables, or by 'projecting' the high dimensional data into a lower dimensional space. Methods for variable selection and projection are discussed in the next section.

There are a number of alternatives to two and three dimensional scatterplots where more than three variables are represented in some way on the same plot. Examples of such methods are Chernoff faces, Andrews plots and weather vane plots. Details of these and other display techniques are given in [Tufte, 1983] and [Everitt and Dunn, 1991]

Reducing the Dimensionality for Classification: Feature Extraction

Developing a system for classification can be described as a two stage process:

- extraction of features which can be used to describe the data
- developing a classification algorithm for the features, i.e. the discriminant functions.

These two stages may overlap considerably and the process will normally be an iterative one, since the choice of features will depend on the success of the classification algorithm and vice-versa. Feature extraction and classification is a process by which the original data are successively refined and reduced until the optimum number of features is selected for classification. The first stage of this process is to extract measurements from the data. These may be the original data, or they may be processed or transformed in some way, for example to remove artefacts or noise or to make them more amenable to the extraction of relevant features. The second stage consists of selecting the subset or combination of these features which give best discrimination. The ultimate stage consists of developing a rule for classifying the data. This can be thought of as an extension of the feature extraction stage; the 'features' from this stage are the values from the resulting discriminant rule. In the case of LDA the number of features extracted from this last stage will be the same as the number of discriminant functions used.

Feature extraction for classification is concerned with finding the best patterns to discriminate and classify data. In the case of MR spectra this means choosing appropriate features to represent the spectra, and then finding which combination or subset of these features provides the best discrimination.

Broadly speaking, features are any extractable measurements that can be used for classification; their choice may involve pre-processing the data very little, for 'low-level' features, or may necessitate a large amount of pre-processing for 'high-level' features. This choice will involve a trade-off between the computational feasibility of using low-level features compared with the inevitable loss of information involved with the extra processing for higher level features [Schalkoff, 1992].

The terms feature extraction and selection are somewhat ambiguous and are often used synonymously, especially in literature on statistical pattern recognition e.g. [Fukunaga, 1990]. In this thesis the term feature extraction is used to describe the whole process of extracting suitable measurements from the 'raw' data through to developing the discriminant rule. Feature selection is used to describe the process of selecting which subset of measurements to use in the classification algorithm and feature reduction is used to describe reduction of features by combining the original variables into a smaller set of new variables. Another somewhat ambiguous term is pre-processing. In this thesis this term is used for any processing of the Fourier transformed datapoints prior to extraction of measurements from the spectrum.

While the number of initial features may be very large, the underlying dimensionality of the data, that is the intrinsic dimensionality [Fukunaga, 1990], may be quite small. Thus it is generally possible to partition the feature space into subspaces of signal and noise. The goal of feature extraction is to eliminate a significant number of dimensions so as to encourage a parsimonious representation of the underlying structure [Scott, 1992]. The following section describes methods that can be used to give this parsimonious representation. This section is in two parts: the first part describes methods which reduce the variables by combining the originals into a smaller number of new ones, the second part discusses methods for selecting a subset of features for classification.

Before discussing these methods it should be pointed out that there are two potential problems which have to be considered when there are a large number of variables to choose from and the number of samples is relatively small. The first problem is that it will always be possible to construct a boundary that will completely separate the classes in the training set if the number of samples is equal to or less than the number of variables. This boundary, however, may completely fail to separate the test set. This problem is known as over-fitting. Another related problem which may lead to poor classifier performance of new data, and also incorrect assumptions about the power of discrimination of the selected variables is selection bias. Selection bias occurs when the subset of variables is not chosen independently of the data used to test the discriminant rule. Ideally the discriminant rule should be developed completely independently of the test data. However, in practice the number of samples is often not large enough to do this and methods such as leave-one-out must be used for validation. Selection bias is a problem because the variables that are chosen to separate one particular sample set of data may not be the best ones for discriminating between the groups in the population as a whole.

Selection bias is often ignored in the literature on pattern recognition. It is covered in detail in [Miller, 1990]. Since its effects are easy to overlook and may only be discovered when attempting to classify new datasets, it is important to be aware of this problem.

Feature Reduction

In this section methods are discussed in which the feature reduction is achieved by combining the original variables into a number of new ones. With the exception of SELECT, all of these methods are 'unsupervised' in that the combination is carried out without using information of the class of the samples. The first method discussed in this section, principal components analysis (PCA), is very widely used as a dimensionality reducing technique, both for data display and classification. The second, factor analysis, which is similar to PCA in some respects, is a useful technique for

reducing dimensionality when the data can be explained by a small number of underlying factors. The third method, the wavelet transform, has proved to be an extremely successful method of compressing many types of data and is a useful method for feature extraction for 'peaky' data such as spectra. The fourth method SELECT is a combination of feature selection and feature reduction, the aim being to provide uncorrelated features for classification. Finally I discuss two methods, non-linear mapping, and projection pursuit that may be useful for reducing the dimensionality for data display.

Principal Component Analysis (PCA) One of the most simple and commonly used statistical method for reduction of dimensionality is principal component analysis (PCA). PCA operates by transforming the original variables into a new set of uncorrelated variables called principal components (PC's). These new variables are linear combinations of the originals derived in decreasing order of importance so, for example, the first PC accounts for as much as possible of the variation in the original data. If the original variables are highly correlated (effectively 'saying the same thing') the first few PC's will account for most of the variation and the remaining PC's can be discarded with little loss of information. Ideally the first few components will be intuitively meaningful, will help us understand the data better, and will be useful in subsequent analyses where we can operate with a smaller number of variables. In practice it is not always easy to give 'labels' to the components and their main use is to reduce the dimensionality of the data in order to simplify later analyses [Chatfield and Collins, 1980] [Massart *et al.*, 1988] [Howells *et al.*, 1992b]. Essentially PCA uses the covariance matrix derived from the co-ordinate system of the original data, x, to construct a new co-ordinate system z. The axes of the new co-ordinate system are orthogonal and lie along the directions of maximum variance in the original data. Both z and x have dimension n. The components of z can be expressed as

$$z_n = w_{n1}x_1 + w_{n2}x_2 \ldots + w_{nN}x_n \tag{3.6}$$

where w_{n1} are constants. The nth component of z, z_n is called the nth principal component.

The principal components are calculated using the covariance matrix, i.e.

$$z = A^t x \tag{3.7}$$

where A^t is the transpose of A, which is composed of the eigenvectors, w_{n1} of the sample covariance matrix, S. The magnitude of the nth largest eigenvalue, λ_n reflects its relative contribution to the variance of z. It is common to calculate the principal components after the original variables have been standardised to have unit variance which is equivalent to using the correlation matrix, rather than the covariance matrix in equation 3.7.

Normally it is possible to compute n principal components from n variables. However, if some of the original variables are linearly dependent, or if the number of variables exceeds the number of samples (i.e. the covariance matrix has rank less than n) some of the eigenvalues (and therefore the PC's) will be zero. For most applications, those which express a certain percentage of the variation, for example 90%, will be chosen and the rest discarded. When the original variables are highly linearly correlated with one another it may be possible to discard most of the principal components with very little loss of information. However if the original variables are nearly uncorrelated the PCA will simply find components that are close to the original variables, but arranged in decreasing order of variance and nothing will have been gained.

PCA is sometimes called an unsupervised pattern recognition method since the derivation of the components uses no prior knowledge of the class of the samples. It has the advantage over other methods of feature extraction based on class differences, for example correlation methods, in that the components can be selected purely on the basis of the variation they explain again using

no class knowledge. This means that no selection bias will be introduced if the test set is used in the feature selection process, as will normally be the case when the leave-one-out method is used.

However, a disadvantage of selecting the PC's on the basis of the variance they explain is that they may not necessarily provide the best features for classification. It is quite common in practice to find that the vector which is most highly correlated with class is one corresponding to one of the smaller eigenvalues [Miller, 1990]. Another disadvantage of this technique is that it is sample dependent and can be unstable if there are a large number of variables compared with samples. Subsequently the inclusion of one or two extra samples may completely change the composition of the PC's. [Chatfield and Collins, 1980] gives a comprehensive discussion of the benefits and drawbacks of PCA.

As mentioned above the goal of feature extraction is to eliminate a significant number of dimensions by partitioning the feature space into subspaces of signal and noise. With PCA the hope is that most of the variance in the data set will be explained by the signal and little by the noise. If this is the case, the partitioning can be achieved by selecting the PC's which account for most of this variance.

PCA is a useful method for reducing the dimensionality for both data display and classification, and is a very widely used method of dimensionality reduction. It is suitable for spectral data because it often works well when the underlying dimensionality of a large feature space is small. However it is only appropriate if the underlying relationships between the variables are linear, and most of the variance in the data is accounted for by the signal rather than the noise.

PCA has been widely reported for reducing the dimensionality of MRS data and the use of this method has been reported by a number of groups using PR methods to analyse medical MRS data. These studies have used PCA either for reducing the dimensionality for display, as a pre-processing step before further analysis, or for investigating biochemical features of the data. Howells and co-workers use PCA as a preprocessing technique to reduce the dimensionality of the original data set. This group developed an automated technique whereby they extract values from the spectra by splitting the spectra into a number of equal intervals, and recording the highest value from each interval. In [Howells *et al.*, 1992b] the original 16,000 datapoints were reduced to 180 variables by the digitisation procedure and PCA was then used to further reduce the number of features to 15 PC's. These 15 PC's, which accounted for 95% of the variation in the data set, were used as input to a neural network classifier and clustering program with good results. This technique was also used in more recent studies, for example [Howells *et al.*, 1993b].

Somorjai also used PCA to reduce the dimensionality of the original data set. In a study using spectra obtained *in vitro* of thyroid neoplasms the spectra were first divided into two sub regions of 170 and 400 points respectively and then PCA was applied directly to the datapoints from these sub-regions. Ten PC's which accounted for 97% of the variance for the date set were retained and successfully used to discriminate between spectra from normal and cancerous tissue.

Other groups report using PCA as a dimensionality reducing technique for data display, most notably the group including Nicholson, Lindon et al, who have applied PR techniques to a number of studies of MRS data of body fluids, for example a study of ^{1}H spectra of urine from rats [Gartland *et al.*, 1991]. In this study the rats were exposed to a number of different toxins. The aim was to characterise the spectra according to the biochemical effects from the different toxins. The original feature vector consisted of signal intensities for 17 metabolites and this was reduced to two or three features using PCA. The data analysis was then extended to obtaining spectra taken at three time points after exposure to the toxins. The PC's from the two sets of data were then used to display and categorise the spectra. In another study by the same group [Holmes *et al.*, 1994] PCA analysis was used to classify ^{1}H spectra of human urine. In this study, 'descriptors' were

automatically extracted from each spectrum by segmenting it into consecutive non-overlapped regions and integrating the signal intensity in each region.

In another reported study [Confort-Gouny *et al.*, 1993] used PCA to investigate correlations between different metabolite measurements from ^{1}H spectra of various brain diseases.

Factor Analysis Factor analysis is concerned with whether the covariance or correlations between a set of variables $X = [x_1, ...x_n]$ can be 'explained' in terms of a smaller number of unobservable latent variables or factors $f_1, \ldots f_k$ where $k < n$. The factor model is given by

$$\begin{aligned} x_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \ldots + \lambda_{1k} f_k + u_1 \\ x_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \ldots + \lambda_{2k} f_k + u_2 \\ &\cdot \\ &\cdot \\ x_n &= \lambda_{n1} f_1 + \lambda_{n2} f_2 + \ldots + \lambda_{nk} f_k + u_n \end{aligned} \tag{3.8}$$

where the λs are the weights or factor loadings and u_j is the residual variation specific to the j^{th} variable. There are a number of methods for determining and choosing the required number of factors. The techniques for finding the factors are somewhat complicated and the interested reader is referred to [Malinowski and Howery, 1980].

PCA and factor analysis are often confused and one technique is often applied erroneously when the other is more appropriate. It is therefore important to appreciate the different aims of the two techniques. The vital difference is that PC's are the optimal entities for expressing variance in the data while factors are appropriate when trying to explain covariance in a multivariate system. The object of PCA is to find a lower-dimensional representation that accounts for the variance of the features. The object of factor analysis is to find a lower-dimensional representation that accounts for the correlations among the features [Duda and Hart, 1973]. Although factor analysis has been much criticised in the literature by theoretical statisticians [Chatfield and Collins, 1980] [Everitt and Dunn, 1991] it can be very useful in fields where combinations of variables can be expressed as meaningful attributes of the data [Malinowski and Howery, 1980]. Factor analysis is useful for exploratory data analysis, in particular for helping the analyst understand the underlying nature of the data better. If the identification of factors is successful, it can be used as a dimensionality reduction technique for both data display and classification but, unlike the other methods discussed in this chapter the expertise of a biochemist will be required in order to make best use of this technique.

Factor Analysis can be very useful technique in analytical chemistry for uncovering hidden characteristics of chemical mixtures. For example it has been used successfully for MRS data to predict the shifts of simple solutes in solution. [Malinowski and Howery, 1980] give an extensive review of the applications of factor analysis in analytical chemistry. Howells et al. have used factor analysis to help interpret ^{1}H MRS spectra of extracts of tumours and normal tissue from rats. They used an approach called target factor analysis which allows for physically significant models of the data to be developed and factors to be tested individually. Using this technique it was possible to determine which metabolites had the greatest influence in the data and were responsible for the separation into groups [Howells *et al.*, 1992b].

The Discrete Wavelet Transform Wavelet theory is a relatively new branch of mathematics which has rapidly found applications in a number of wide-ranging disciplines including physics, numerical analysis, signal processing, probability and statistics. The usefulness of the wavelet transform

is due to the fact that it can be used to approximate functions/signals according to scale resolution using a set of basis functions called wavelets. Wavelets allow a representation of the original function in which both scale and spatial information are retained. Many functions can be approximated very closely using only a small number of wavelet coefficients. The wavelet transform may also be used to represent economically, localised features of interest in a signal, which makes it an ideal candidate for extraction of features for classifying spectra. The wavelet transform is not strictly a method of statistical pattern recognition, rather it is a pre-processing method which allows the data to be expressed more succinctly. However, since this is the aim of feature extraction it is appropriate to include it in this section.

The use of the wavelet transform for feature extraction can be described in a number of ways. It can be used as a filtering technique for removing the high frequency components from the data, or used as a method for representing shape information in a succinct way. Alternatively, it has excellent data compression properties.

The Discrete Wavelet Transform (DWT) transforms a data vector of length n into another vector of length n wavelet coefficients using a set of n orthogonal basis functions called wavelets. Each wavelet coefficient is calculated by taking the dot product of the data vector with one of the basis functions. The set of basis function is derived from a single function (often called the 'mother wavelet') by a series of dilations and translations. The DWT is similar to the Fourier transform in some respects but, unlike the sine and cosine basis functions of the Fourier transform, wavelets are localised in space as well as in scale.

In order to discuss the advantages of the DWT it is useful to compare it with the windowed or short time Fourier transform, since this has been one of the most popular classical techniques for pre-processing data with localised features [Daubechies, 1992]. The Fourier transform is a good method for representing data where the small scale (i.e. high frequency) features represent the detail or noise originating in the signal or function, and the large scale (low frequency) features represent the basic shapes. However it has the disadvantage that the frequency information obtained from a Fourier transform is global, because its basis functions are sine and cosine functions. This is not satisfactory when localised features are required. This problem may be partially overcome by using short-time or windowed Fourier transforms whereby the signal to be analysed is multiplied with a window function before computing its Fourier transform. However this method has the problem that a window of fixed size in the original domain is accompanied by a fixed sized window in the Fourier domain. What is really needed is a long window to analyse large scale components and a narrow one to detect the small scale features [Wunsch and Laine, 1995]. This is exactly what is provided by the wavelet transform.

The set of basis functions is obtained from the mother wavelet $g_{basic}(t)$ by dilations controlled by the variable a, and translations controlled by the variable b, according to the equation

$$g_{a,b}(t) = \frac{1}{\sqrt{a}} g_{basic}\left(\frac{t-b}{a}\right) \tag{3.9}$$

A function g said to be a wavelet if it satisfies the following admissibility condition needed to obtain the inverse of the wavelet transform

$$\int |\hat{g}(\omega)|^2 \frac{d\omega}{|\omega|} < \infty \tag{3.10}$$

where $\hat{g}(\omega)$ denotes the Fourier transform of $g(t)$.

Also the function should have finite energy i.e.

$$\int |g(t)|^2 dt < \infty \tag{3.11}$$

$$and\ \hat{g}_{basic}(\omega) = 0\ for\ \omega < 0 \tag{3.12}$$

[Bos and Vrielink, 1994]

There are many types of wavelet transform, and the discussion in this chapter is limited to those in which the set of basis functions is orthonormal. The wavelet transform of a signal, $f(x)$ at scale 2^j is defined by the dot product:

$$w_{2^j,b} = \sum_{i=1}^{n} f(i) g_{2^j,b}(i) \tag{3.13}$$

[Mallat and Zhong, 1992] where $w_{2^j,b}$ is the wavelet coefficient at scale j and position b. The wavelet transform can described in a number of ways. An elegant way is Mallat's multi-resolution analysis [Mallat, 1989]. This describes a wavelet transform as a process of representing a function at different levels of approximation. The first level approximates the function very generally by projecting it onto a space spanned by two basis functions of large scale, the second level will be a little less general (using 4 basis functions) and so on. The last level will completely represent the signal and the inverse transform of this representation will yield the original signal. Mathematical details are given in [Mallat, 1989] [Daubechies, 1992]. The DWT can thus be used to partition the feature vector into a sequence of subspaces, each subspace representing a different scale level.

The DWT algorithm operates by transforming the original feature vector into a new vector which is filled sequentially with the wavelet coefficients of the different scales. Each scale corresponds to a different dilation of the mother wavelet. Furthermore, within a scale, each wavelet corresponds to a different translation of the mother wavelet. The lower numbered coefficients represent the large scale (low frequency) features in the original feature vector and the higher numbered coefficients represent small scale features. Thus for a vector of length 2^j the coefficients are ordered into j scale levels – each scale level representing the data vector at a certain resolution. Scale level 1 is represented by wavelet coefficients 1 and 2. Scale level i, where $i = 2 \ldots j$, is represented by wavelet coefficients numbered from $2^{i-1} + 1$ to 2^i. From this it can be seen that the number of coefficients which represent a scale level increase by a factor of two as the scale of the features decreases. This has the desired effect of adapting the window size to the scale of the features. This means that the number of coefficients used to represent high frequency information is substantially more than those representing low scale information. If these represent just noise these can be discarded with little loss of information. Unlike sines and cosines which define a unique Fourier transform there is not one single unique set of wavelets: in fact, there are infinitely many possible sets. The exact choice of mother wavelet involves a trade-off between their degree of smoothness and their degree of localisation. A popular series of wavelets is the Daubechies series which we have used for this study [Press *et al.*, 1992]. Figure 3.1 shows some of 512 basis functions used in this study, derived from the Daubechies 20 mother wavelet [Press *et al.*, 1992].

So far, reported applications of the wavelet transform for feature extraction for classification have been relatively few. Wunsch et al. [Wunsch and Laine, 1995] used the wavelet transform to extract useful features (which they call wavelet descriptors) for classifying handwritten characters, making use of the fact that shapes that 'look alike' often have similar low frequency components. In this study, the wavelet transform provided a good alternative to using the windowed Fourier transform for extracting features for classification. The two feature extraction techniques were compared using a neural network classifier on over 600 samples of hand printed characters. The results showed that the wavelet descriptors provided better classification results than the Fourier descriptors. [Saito, 1994] discusses the use of the wavelet transform for localised feature extraction

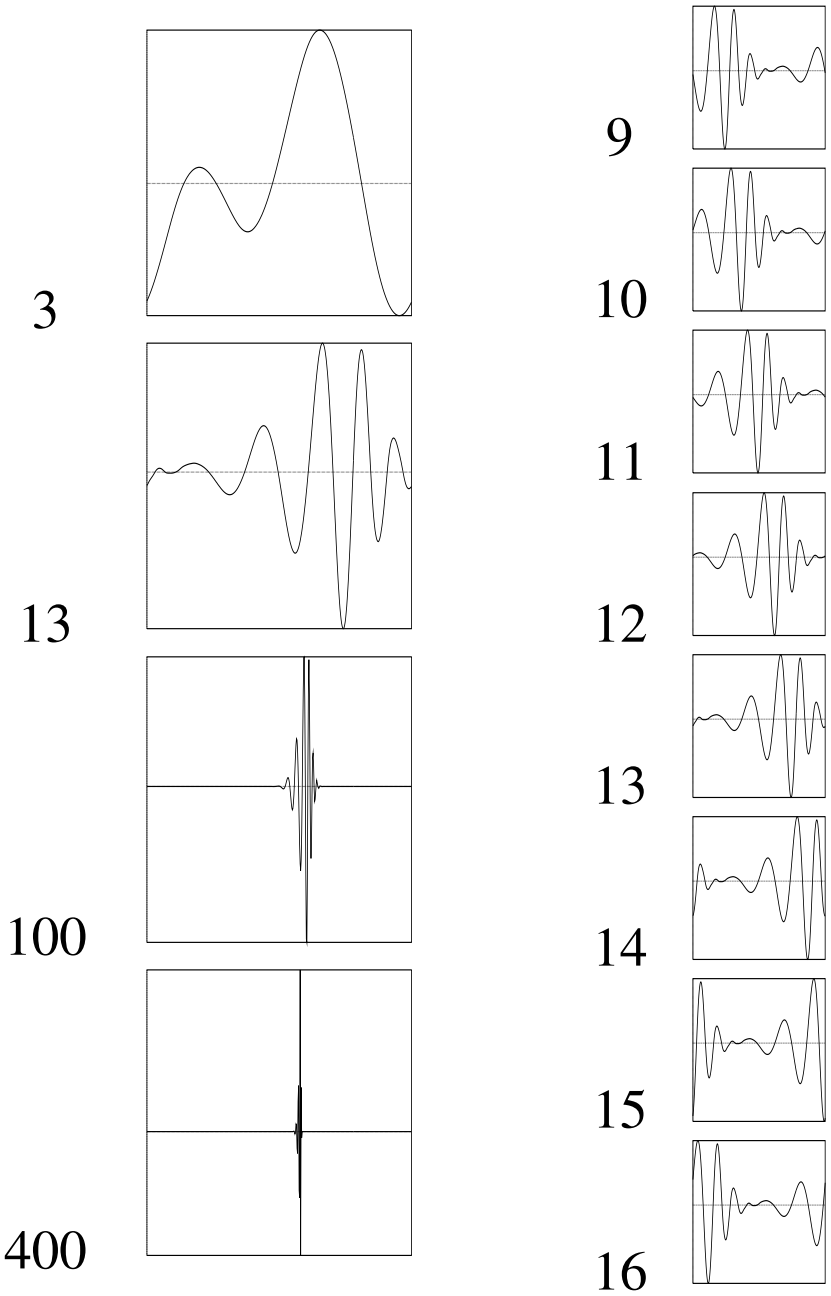

Figure 3.1. Some of the Daubechies 20 wavelets used in this study. The left hand column shows the wavelets from four different scales. The right hand column shows a collection of wavelets which together represent scale level 4. In both columns, the number to the left of the wavelet indicates the number of the wavelet.

from geophysical data. The first (and as far as I know the only) reported use of the DWT for classifying spectral data is by [Bos and Vrielink, 1994], who used wavelets to classify IR spectra of different compounds. In common with MR spectra, the relevant information in the IR spectra is contained in the position and shape of the absorption peaks. This study showed that the wavelet transform, due to its localisation both in position and scale, can extract this information in a concise form and thus can be used to extract the salient feature from an IR spectrum effectively.

Comparison of DWT with PCA Although the algorithms for computing these transform are fundamentally different, they can both be used to achieve the same aim, that is to partition the feature space into subspaces of signal and noise. Both transforms operate by combining the n features into n new features. However, whereas the PC's are uncorrelated the wavelet coefficients are not. Another major differences between PCA and the wavelet transform is that the first few wavelet coefficients, i.e. the low frequency components represent the variation within the sample, whereas the first few PC's will represent the variation between the samples.

SELECT SELECT is a method of feature extraction proposed by Kowalski [Kowalski and Bender, 1976]. Although its use has not been widely reported in the literature, it is very suitable for spectral data and provides an alternative to PCA which incorporates knowledge of class differences. As in PCA the original variables are combined to provide new features which are orthogonal to each other. However, unlike PCA these new features retain some of their original identity and are thus easier to relate back to the original data. The procedure begins by selecting the most important discriminatory feature using an appropriate feature selection method and making this the first new feature. The remaining features are then de-correlated with the one that has been selected and the process is repeated until the required number of new features have been obtained. In common with PCA the feature vectors are rotated in the pattern space. However, in this case the new features are selected on the basis of the differences between the classes, rather on the amount of variation that they express. In this method the first component is just the best discriminatory variable – chosen by a feature selection method such as weighted variance – the second feature is a linear combination of this feature and the second 'best' feature, the third feature is a combination of the first, second and third selected features and so on. [Kowalski and Bender, 1976] and [Sjostrom and Kowalski, 1979] compared SELECT with PCA as a method of feature selection on a number of data sets, including a set of ^{13}C spectra of exo- and endo-substituted norbonanes and found the two methods gave comparable results. The advantage of SELECT over PCA is that the features may be more easily related back to the original measurements.

Projection Pursuit is a method first suggested by Friedman and Tukey [Friedman and Tukey, 1974] which seeks to linearly project the data onto subspaces which produce the most 'interesting' configurations of the data set. Principal components analysis can be thought of as a projection pursuit method for which the interesting configurations are those with high variances. For many applications PCA gives excellent results; however high variance is not the only relevant criterion of the genuine importance of structure displayed in a projection of the data. Projection pursuit uses other criteria of importance and then uses a numerical optimisation technique to find the projection of the most interest [Silverman, 1986] [Glover and Hopke, 1992].

Non-linear mapping Non-linear mapping (NLM), also called multi-dimensional scaling, is a dimensionality reducing technique in which the criterion is to minimise the differences between the inter-point distances in the higher dimensional space and inter-point distances in the new lower dimensional space [Massart *et al.*, 1988]. If the original distances between objects are denoted by d_{ij} (for objects i and j) and the new distances (in a two dimensional space) by d^*_{ij}, one searches for those d^*_{ij} for which the differences with the d_{ij} are as small as possible. Many algorithms have been proposed for multidimensional scaling. Several of these are based on the minimisation of the

so-called mapping error E [Massart *et al.*, 1988] where

$$E = \sum_{i<j} \frac{(d_{ij} - d^*_{ij})^2}{d_{ij}} \tag{3.14}$$

Non-linear mapping methods can be used to help determine the underlying dimensionality of the data but are generally used to find the best two or three dimensional representations for data display. Non-linear mapping was applied by [Gartland *et al.*, 1991]. In this study of ^{1}H spectra of urine from rats PCA provided consistently better results than NLM in terms of discrimination of toxicity types. The same group have more recently used this technique in a study of MRS data of human urine [Holmes *et al.*, 1994].

Feature Selection Methods

Feature selection is concerned with choosing the best variables to use in the discriminant function algorithm. This has the advantage over feature reduction methods such as PCA in that the original identity of the variables is maintained, which is important if we are trying to determine the differences between groups. However, most of the traditional methods of feature selection assume a small number of variables and may not be of much help for data such as spectra which have a very large number of variables.

Apart from the fact that feature selection reduces the number of variables for classification it has another advantage in that it can be used to reduce the number of measurements needed in the original experiment, which can be an important consideration. Once important features for discrimination are found the original experiment may be adapted so that just these features are measured.

Methods for subset selection When the number of variables is not too large, standard subset selection techniques may be used. Finding the best subset of m variables out of n may be carried out by evaluating a criterion of class separability for all possible combinations of the m variables. However, since calculation of $\binom{n}{m}$ combinations becomes prohibitively expensive for even fairly small values of n and m, it is necessary to use procedures which avoid exhaustive search. Examples of such procedures are backward and forward stepwise selections and branch and bound. Details of these methods are found in [Fukunaga, 1990]. Most statistical software packages incorporate subset selection techniques into the discriminant programs, and often the criteria of separability will be the same, or closely related to those used for determining the discriminant function.

Methods for selecting individual variables If there are large numbers of variables it is generally preferable to preselect a few of the variables prior to subset selection. This will be the case when using spectral datapoints as the measurements since many of these represent noise, rather than natural features of the data. Good features must satisfy the following criteria: intraclass variance must be small which means that features derived from different samples of the same pattern class should be close and interclass separation should be large, i.e. features derived from samples of different classes should differ significantly [Wunsch and Laine, 1995].

Features whose means differ widely between classes but who have small intraclass variance can be selected using this equation

$$\frac{\bar{x}_{Kj} - \bar{x}_{Lj}}{\sqrt{s^2_{Kj} + s^2_{Lj}}} \tag{3.15}$$

where $\bar{x}_{Kj}$ is the mean of variable, x_j for class K and s^2_{Kj} is the variance for the same variable of this class [Massart *et al.*, 1988].

Another method called variance weighting permits weight to be given to the variables on the basis of their power to discriminate between the training sets. These weights are measures of between class variance to within class variance for the groups. For two classes K and L the weights are obtained by using the equation

$$w_j = \frac{\frac{n_K \cdot n_L}{n^2} \sum_{k=1}^{n_k} \sum_{l=1}^{n_l} (x_{kj} - x_{lj})^2}{\frac{n_K}{n} \sum_{k=1}^{n_K} \sum_{k'=1}^{n_K} (x_{kj} - x_{k'j})^2 + \frac{n_L}{n} \sum_{l=1}^{n_L} \sum_{l'=1}^{n_L} (x_{lj} - x_{l'j})^2} \tag{3.16}$$

[Massart *et al.*, 1988]

Another possible strategy for selecting individual variables, suggested by [Massart *et al.*, 1988] is to perform a cluster analysis of the variables over all the samples to see if a grouping of the variables indicating some sort of similarity in behaviour over the data set can be found. If it can and if the variables cluster into a small number of groups, then one variable can be selected from each group.

Correlation Methods If there are only two classes, or if the classes can be ordered, a simple method of finding which variables vary most with the class of the objects is to calculate correlation coefficients. The most commonly used measure of correlation is Pearson's correlation coefficient – also known as the product-moment coefficient.

$$r = \frac{\sum_{i=1}^{i=n} (x_i - \overline{x}_i)(y_i - \overline{y}_i)}{\sqrt{\{\sum_{i=1}^{i=n} (x_i - \overline{x}_i)^2 \sum_{i=1}^{i=n} (y_i - \overline{y}_i)^2\}}} \tag{3.17}$$

x and y are the two variables and $\overline{x}$ and $\overline{y}$ are their respective means. This coefficient gives a measure of association between numerical variables. Since class is a qualitative rather than a numerical variable it is first necessary to assign an appropriate numerical value to each class (for example 0 and 1 if there are two classes). Pearson coefficients are appropriate if there are only two classes, or one can be sure that the classes are equally spaced, since the variables must represent measurements from an equal interval scale. Otherwise Spearman coefficients are more appropriate.

Spearman coefficients give a measure of association of variables that can be ranked. This coefficient is calculated by replacing x_i and y_i in the above equation by their rank ordering. If the x_i and y_i are replaced by the first n integers then equation 3.17 becomes

$$r = 1 - \frac{6 \sum_{i=1}^{i=n} d_i^2}{n^3 - n} \tag{3.18}$$

where d_i is the difference in ranks between x_i and y_i for individual i. Note that a correction for ties is necessary if ties occur in one or both of the series of ranks. Full details of this and other measures of correlation are given in [Krzanowski, 1988] [Siegel and Castellan, 1988].

The above correlation coefficients always take values between -1 and +1, with +1 or -1 representing a perfect linear association (positive or negative respectively) and 0 none. Both methods measure linear association between variables and thus do not account for non-linear relationships. However, this limitation of the technique does not apply to investigating the relationship between class and other variables when there are only two classes.

Tests for significance of Pearson's correlations are based on the assumption that both variables are normally distributed, whereas the tests for significance for Spearman's correlation use no such assumptions. Because of this the Spearman methods is often called nonparametric correlation. Since a high absolute value of the correlation coefficient indicates a strong association

between the variable and the class of the object these coefficients may be used to select variables for discrimination.

Apart from calculating these coefficients between individual variables with class, it is very useful to calculate a correlation matrix for the complete set of variables, prior to choosing which feature extraction methods to use. This is very useful both as an exploratory tool and also to investigate whether methods that depend on the correlations between variables, for example PCA might be appropriate. One of the features of *in vivo* MR spectra is that the peaks are relatively broad, thus the number of values representing a peak may be spread over a number of datapoints and these values will be highly correlated with one another. The same applies to datapoints from coupled peaks (see chapter 2). Thus a strategy would be to choose one datapoint, for example the central datapoint or that most highly correlated with the class from each highly intercorrelated region. This strategy was used in this research and is discussed in more detail in the next chapter.

The correlation coefficient is often used to select wavelengths for calibration of infrared spectra. [Wu *et al.*, 1995] compared the use of correlation coefficients with the two feature selection methods described by equations 3.15 and 3.16 for selecting wavelengths for classifying near-infrared spectra of drug tablets. This study found that similar features were selected by all three methods. To my knowledge neither correlation methods, nor any of the other feature selection methods mentioned in this section, have been used for selecting variables for medical MRS data analysis.

3.6 Previous Work

The need for a multivariate approach to spectra analysis is now widely acknowledged by researchers studying medical MRS data and studies by several groups, for example Howells et al. Nicholson et al. and Somorjai et al. have demonstrated the usefulness of the pattern recognition approach, both for helping build up a metabolic profile of the tissue, and also for discovering the best way of discriminating between different classes of spectra. However, these methods are currently used by only a few centres analysing medical MRS data, and most analysts still use semi-manual methods for obtaining measurements from the spectra, and univariate methods for subsequent interpretation.

At present there are only a few research groups specialising in statistical pattern recognition analysis of medical MRS data. The major 'players in the field' seem to be: Howells and co-workers who have applied many of the chemometric techniques mentioned in this chapter to both *in vivo* and *in vitro* data, using neural networks and cluster analysis for classification, Lindon, Nicholson et al. who specialise in PR for classifying spectra of body fluids, with particular emphasis on data display methods, and the Winnipeg group (Somorjai and co-workers), who have developed a 'computerised consensus diagnosis' method which uses a number of different techniques to classify biopsy data. Reference to these studies are made in the relevant sections.

Most of the reported studies using statistical PR are concerned with data that has been acquired *in vitro*. For example Nicholson et al. have carried out studies on body fluids such as urine and cerebrospinal fluids. Others have carried out studies of biopsy data, the Winnipeg group have used PR techniques to analyse biopsies of thyroid neoplasms, cervical dysplasia and brain tumours for example. Howells and co-workers have applied PR analysis to biopsies of rat tumours.

However, there is an increasing interest in using PR analysis of *in vivo* data especially for classifying tumours. Several recent studies have demonstrated good results classifying proton brain tumours, using statistical techniques, for example [Preul *et al.*, 1994] [Hagberg *et al.*, 1995]. These studies, however, are based on using values that are extracted after editing the spectra individually

and quantifying the peaks semi-manually.

Each of the three groups mentioned above have developed techniques for automatically extracting values from the spectrum (see section 3.5.3). However, as far as I am aware, Howells et al. are the only group to have applied an automated feature extraction technique to *vivo* data [Howells *et al.*, 1993b].

The previous work shows that either linear discriminant analysis, neural networks or data display techniques, can be used successfully to automatically classify MR spectra obtained *in vitro* for tissues displaying different types of pathology. A few studies show that this approach may also prove successful for classifying *in vivo* data. However, as yet no fully automated technique has been developed for this task. More powerful techniques for extracting features automatically from the spectrum are therefore required. Investigating such methods is the main purpose of this research.

3.7 PR Methods Used in this Research

This research extends previous work on *in vivo* MRS data by exploring methods for fully automating discrimination between spectra of different tissue types or classes. While studies using PR techniques have been used for exploring relationships between features of spectral data, and also for classifying MRS data, none have been explicitly aimed at exploring methods for fully automated feature extraction for classification, using no prior knowledge of the relative importance of the resonance frequencies.

In this research the method of choice for discrimination was linear discriminant analysis. There were two main reasons for this. The first was driven by necessity since the sample to variable ratio of the test data made many of the other methods in this chapter impracticable. The second was that it is generally preferable to try the simplest approach first, only going on to more sophisticated methods if this doesn't work. This approach was taken both for feature extraction and classification. A potential problem of using more complicated techniques is that they may obscure the results making subsequent interpretation much more difficult. Another problem is that the more parameters that need to be estimated the more the analysis will depend on the test data and also possibly the subjective judgements of the operator.

The approach taken for feature extraction was to select features purely on the basis of their power to discriminate between different types of spectra, using no prior knowledge of biochemistry. These features were chosen using a combination of data display and statistical techniques. First the spectra were plotted and examined individually for obvious differences between the classes. Secondly a 'mean spectrum' for each class was created and displayed on the same plot, in order to identify which regions or datapoints in the spectrum might provide the best discrimination. On the basis of this preliminary investigation three types of features were selected

1 peak heights

2 spectral datapoints

3 wavelet coefficients

A combination of feature reduction and feature selection methods were then used to reduce the dimensionality of the resulting feature vectors. These included PCA for feature reduction, and correlation methods for selecting individual features. The methodology used is described in detail in the next chapter.

3.8 Summary

This chapter gives a review of the statistical pattern recognition methods that may be useful for the analysis of clinical MRS data, and introduces the methods that were used in this research. In the first section the two other approaches to pattern recognition analysis, that is the neural and structural approaches, and some basic terminology, are discussed briefly. The main part of the chapter is devoted to providing the motivation for using PR methods for analysing MRS data together with a description of some of the methods that may be appropriate for describing and classifying MRS data.

Spectral data can be described either by explicit peak measurements, or by other features such as spectral shapes, or linear combinations of the spectral datapoints. Most of the methods described here are appropriate for either choice of features. However, since the aim of this research is to extract features for classification using the whole spectrum as the initial feature set, more attention was paid to methods for feature extraction which can cope with a large number of datapoints compared with the number of samples. While some of the methods that were discussed are general methods applicable to most types of data, also included are methods particularly suitable for spectroscopic data, such as factor analysis, the DWT and SELECT.

The final part of this chapter gives a report of the 'state of the art' of PR analysis of MRS data. This section does not provide a comprehensive literature review since it seemed more appropriate to refer to relevant reported studies within the sections on the methods instead of adding them at a separate section at the end. Rather, its purpose is to put this work in context with previous studies. As noted at the beginning of the chapter, while much effort is being applied to the development of improved techniques for the acquisition and processing of MRS data, relatively little effort has been applied to the development of methods for the interpretation of medical spectra. This is reflected in the relatively small number of reports of PR analysis of MR spectra which is particularly true for in *vivo* data. Since it is as a tool for *vivo* analysis that spectroscopy really comes into its own I hope that this research, which develops a prototype system for automatically classifying *in vivo* spectra may in some way to help to redress this imbalance.

Chapter 4

Developing an Automated System to Classify Spectra

4.1 Introduction

The previous chapters have been devoted to developing the background and presenting the motivation for the work carried out in this research. The purpose of the remainder of this thesis is to demonstrate how some of the methods discussed in the previous chapters can be used to develop a prototype system for classifying *in vivo* MRS data.

The main purpose of this chapter is to discuss in detail the practical problems that are involved in designing a system for automatically classifying MRS data, together with the methods that can be used to for solving these problems. These methods were developed and tested using two sets of *in vivo* data: a set of ^{13}C spectra from human volunteers and a set of ^{31}P spectra obtained from rats. However, since the aim of this study is to develop general methods that can be used to classify different types and classes of *in vivo* spectra, only brief reference is made to these particular data sets in this chapter.

A Pattern Recognition system will normally involve three stages:

- acquisition and pre-processing of the data,
- extraction of features which can be used to describe the data and
- description and classification based on these features.

These stages will normally overlap considerably, particularly the feature extraction and classification stages. The development of a pattern recognition system is normally an iterative process: the data acquisition and feature extraction processes will be modified according to the results obtained from the classifier. In turn the design of the classifier will depend on the type of data and features that are extracted.

This chapter discusses the methods that were developed for each of these stages. Particular attention is paid to the pre-processing and feature extraction stages because they require specialised methods particular to the type of data, and also because they present the greatest challenge for automatically classifying *in vivo* spectra.

While the main emphasis is on providing methods for automated feature extraction, I also bear in mind the ultimate goal which is to provide a system which will facilitate the use of MRS as a clinical tool. The most basic requirements of such a system is that it will take as input an FID signal or spectrum and produce as output a classification for the signal, together with the probability of the subject belonging to that particular class.

However, because the system is intended to be used as a clinical aid, it is also important that it should also provide as much information as possible as to how any decisions were reached. The output of the system should include not only the results of the classification, but also, if possible, the reasons why a subject has been assigned to a particular class. It is unlikely that a classifier will be acceptable to a clinician unless it does so. Thus, although the main priority is to find features that provide the best discrimination, it may be the case that methods which provides slightly less good discrimination may be desirable if these features can be more easily related back to the biochemical nature of the data, particularly if the system is to be used to help in the understanding of the disease process in the tissues being examined. The feature extraction methods investigated in this thesis provide ways of finding the important discriminatory features which do not require any prior knowledge of the biochemical nature of the data, and therefore have the advantage of providing an alternative 'view' of the data from that obtained by the more traditional methods of spectral analysis.

Another important issue to be considered when designing a system to be used to facilitate the interpretation of MRS data is how to present the results of the analysis in a format that is accessible to the clinician. While the development of a user-friendly interface is beyond the scope of this thesis, the issue of how the results may best be presented should be taken into account when considering which methods to use for feature extraction and classification. For example, if the data can be reduced to two or three variables which show good separation of the classes on a scatterplot, it may not be necessary to develop a computer-based classification rule if it is possible to classify the spectra by eye.

The following two chapters are structured as follows: this chapter discusses in detail the development of a general prototype system for classifying MRS data and considers the practical issues that are likely to be encountered at each stage in its development. Chapter 5 then shows how these methods were applied and used to classify successfully two particular sets of *in vivo* data.

4.2 Acquisition and Pre-Processing of the Data

While it is beyond the scope of this thesis to attempt to compensate for unsatisfactory data, it is important to be aware of any problems that might affect the outcome of a pattern classifier. The quality of the data is always of the utmost importance; the term 'garbage in garbage out' is just as relevant for pattern recognition analysis as for any other type of information processing. It is therefore important to ensure that the data is acquired in a consistent manner. This means that the instrumental variables should vary as little as possible between acquisitions, that all the parameters should be the same, and that adjustments to the signal, such as shimming, should be carried out consistently. If the data are acquired using different MR systems, great care will need to be taken to ascertain that the two are compatible; if not the patterns found in the data may be related to the different systems, rather than to natural differences between the groups.

4.2.1 Spectral Processing

The spectrum is obtained from the FID signal by Fourier transformation. A certain amount of pre-processing will be necessary before pattern recognition analysis can be performed. The normal steps for spectral processing are [Gadian, 1982]:

1. store the accumulated free induction decay on magnetic disc or tape
2. remove a possible dc component
3. manipulate the free induction decay by applying an exponential multiplication
4. if desired, zero fill
5. Fourier transform the FID
6. perform the phase correction
7. display and plot the spectrum.

Two further steps are required in order to make the spectra compatible for pattern recognition analysis.

1. align the peaks in the spectrum.
2. normalise (i.e. scale) the spectral datapoints.

An explanation of how these procedures may be carried out is given below.

Processing Prior to Fourier Transformation

As discussed in Chapter 2, there are a number of procedures that are routinely carried out before Fourier transformation in order to enhance the quality of the spectrum. The first procedure is the removal of a possible dc component in the FID, in order to eliminate a possible peak at zero frequency. This procedure is carried out by selecting a number of points at the end of the FID where the NMR signal is small, and subtracting the mean value of these points from each point in the FID. The number of points to be summed will depend on the signal. In this thesis, the calculation was based on the last quarter of the points in the FID. The algorithm used is as follows:

1. Let the FID signal be represented as a data vector of length n, i.e. $X = [x_1 \ldots x_n]$ (n is divisible by 4)
2. Sum the last $\frac{1}{4}$ values of this vector, i.e $[x_{\frac{3n}{4}+1} + \cdots + x_n]$
3. divide this sum by $\frac{n}{4}$ to obtain the mean m of these values
4. subtract m from each value in the vector to provide the adjusted FID $X' = [x_1 - m, \ldots, x_n - m]$

The quality of the spectrum is ultimately determined by the signal and noise that are acquired and subsequently stored within the computer. However the signal-to-noise ratio of the spectrum can be considerably improved by means of appropriate manipulation of the data prior to Fourier transformation. The most commonly used procedure for doing this is apodisation which

entails multiplying the signal point by point by a decaying exponential function, known as a line-broadening filter. This multiplication has the effect of eliminating noise by preferentially lending more weight to the initial part of the free induction decay, where the signal-to-noise ratio is high, than to the latter part where the ratio is much lower [Gadian, 1982]. While the signal-to-noise ratio is considerably enhanced, apodisation does have the effect of increasing the linewidth, and therefore decreasing the spectral resolution. This is particularly undesirable when the peaks are close together or overlapping. The optimal signal-to-noise ratio in the spectrum is achieved when the decaying exponential has the same time constant as the free induction decay (T_2^*) [Gadian, 1995]. The time constant can be determined by the linewidth which is defined as the width of the peak halfway between the baseline and its maximum value. However this can be difficult to measure with spectra obtained *in vivo*, which do not have a clearly defined baseline or peaks, and which also may have peaks of different widths. In practice an estimation is often made of the value of this time constant.

In general, pattern recognition methods for analysing magnetic resonance spectroscopy data will be more tolerant to the presence of noise than more traditional methods of spectral analysis since they combine information from the whole spectrum rather than from just a few peaks. Thus it might be desirable to choose a lower line broadening factor than is normally used for spectral analysis. The approach used in this research was to choose the line broadening factor that maximises measurable differences between spectra from different classes. Details of how this method was applied are given in chapter 5.

Zero-filling involves adding an array of n zeros to the end of each free induction decay of n data points. This has the effect of doubling the number of points in the transformed spectrum. It is possible to zero fill many times. Zero filling will not result in a genuine improvement in the quality of the spectrum, but will be equivalent to interpolating between the genuine points [Gadian, 1982] [Sanders and Hunter, 1993].

Processing after Fourier transformation

Phase correction The aim of phase correction, as mentioned in chapter 2, is to ensure that all the signals in the real part of the transformed spectrum are in the absorption mode. Normally the spectrum is phased interactively by the operator who will adjust knobs on the spectrometer until the spectrum seems to have the correct appearance. Apart from being very time consuming this means that the results will be operator dependent. It is desirable to have a fully automated method for phasing the spectra.

In this thesis an algorithm proposed by [Chen and Kan, 1988] was used to phase the spectra. The phase correction makes use of both real and imaginary parts of a spectrum denoted $F_r(\omega_k)$ and $F_i(\omega_k)$ respectively, where ω_k is the off-resonance frequency for point k. If correctly phased F_r should contain the absorption mode while F_i should contain the dispersion mode only. In order to ensure that this is the case the following combinations are performed to force then into the correct modes:

$$F_r''(\omega_k) = F_r(\omega_k) \cdot cos(\omega_k \cdot t_0 + \phi_0) - F_i(\omega_k) \cdot sin(\omega_k \cdot t_0 + \phi_0) \tag{4.1}$$

$$F_i''(\omega_k) = F_r(\omega_k) \cdot sin(\omega_k \cdot t_0 + \phi_0) + F_i(\omega_k) \cdot cos(\omega_k \cdot t_0 + \phi_0) \tag{4.2}$$

where the double primes denote the final correct mode ϕ_0 is the zeroth order correction and t_0 is the first order correction. The algorithm for performing these correction entails iteratively adjusting the parameter ϕ_0 until the maximum ratio between the highest and lowest values in the spectrum is obtained. The procedure is then repeated if required for t_0. Full details of this algorithm are given in [Chen and Kan, 1988] This algorithm was found to work very successfully on the data studied in this thesis, where only a zeroth order phase correction was found to be necessary.

It is possible to avoiding the problem of incorrect phasing by using the 'absolute' (or magnitude) spectrum

$$\sqrt{F_r(\omega_k)^2 + F_i(\omega_k)^2} \tag{4.3}$$

However, the magnitude spectrum has broader lines than the absorption mode spectrum and therefore increases peak overlap in crowded spectra [de Certaines *et al.*, 1992]. Values from the absolute spectra were investigated in this research but did not give nearly such good classification results as features extracted from the phased absorption mode spectrum. This contrasts with *in vitro* studies carried out by Somorjai and co-workers who report equally good results using either the phased or magnitude spectrum for classifying biopsy data [Dolenko and Somorjai, 1995] [Somorjai *et al.*, 1995a].

Peak Alignment

A problem that must be addressed in order to make the spectra compatible for pattern recognition analysis is the fact that resonance frequencies, and thus peaks, may have different positions in each data vector, due to instrumental factors. This is normally rectified by selecting one peak which is clearly identifiable in each spectrum and using this peak to align the others. This alignment is carried out by finding this peak (i.e. the point with the highest value within the region that this peak appears) and making sure that it has subsequently the same position in each spectrum. This can be done automatically by choosing a suitable value for the new index of this point, finding the index of the highest value in each spectrum and changing this to the new index. In order to do this, it is necessary to discard some of the points at either end of the spectrum, but this is not usually a problem as the range of resonance frequencies in the original spectrum will normally extend beyond the peaks of interest.

For spectra which have clearly identifiable peaks, which do not shift position in the spectrum due to pH or other factors, alignment may be relatively easy. It is not so straightforward, however, when some of the peaks may shift relative to the others due to pH differences as is often the case for ^{31}P data acquired *in vivo*. Another problem which may be encountered is that many of the peaks will be very close together or be combinations of overlapping peaks. In this case it can be difficult to find a single point which can be used to align all the spectra. It is important to take account of pH and other factors when deciding which peak should be the reference for alignment.

The approach used in this thesis was to first ascertain which of the peaks were least affected by shifts due to pH and then to examine the spectra carefully in order to determine which peak seemed to be most clearly identifiable. This peak was then used to align the spectra. The term 'peak' is used in this context to mean the datapoints with the highest value within a predefined sub-region of the spectrum.

The algorithm used to align the peaks in the spectra for peak alignment in this research was as follows:

1. ascertain, by visual inspection, the index, j, of the reference peak in a typical spectrum,
2. ascertain, again by visual inspection of each spectrum, the range of positions, $[j-m, j+m]$, at which this peak occurs for the data set, i.e. how many points (m) either side of the index j to search.
3. decide on the number of datapoints, l_1 and l_2 to be retained after peak alignment either side of this peak. The index for this peak in the new vector will thus be $l_1 + 1$

On the basis of these decisions each spectrum can be aligned automatically using the following steps

1. find the point with the highest value in the range $[x_{j-m}, x_{j+m}]$
2. set the index of this point to $l_1 + 1$,
3. shorten the data vector by removing the points to be discarded at either end of the original data vector.

After alignment, each spectrum can then be inspected visually to ascertain that all the main peaks appeared in the same positions in each spectrum.

To give an example of how the algorithm operates, consider a spectrum which has 1024 datapoints. After the initial investigation a clearly identifiable peak has been observed to always occur within 15 points either side of position 400 in the vector of spectral datapoints. This peak is selected as the reference peak, and the decision is taken to retain 100 points from the original data vector to the left of this point (i.e. the lower indexed points) and 411 to the right of it, so the new vector will have 512 points. The algorithm for automated alignment would then be as follows:

- find the datapoint with highest value in the region of the vector of spectral datapoints indexed from 385 to 415.
- create a new vector with this point at position 101 and adjust the indices of all the other points in the spectrum accordingly,
- discard all points whose new indices are < 1 or > 512.

This peak alignment procedure turned out to be one of the most time-consuming procedures for the ^{31}P data set – see next chapter for details – since it needed several iterations for the procedure to find a suitable subregion and peak. This involved examining each individual spectrum several times. While it was possible to fully automate the alignment process, using the procedure described above, for the data sets examined in this research, this may not always be possible for other data sets. For example, the shifts between the individual spectra may be too large to ensure that the peak chosen for alignment is always the highest within a specified region. If this is the case it will be necessary to first shift the spectra before alignment, possibly by selecting another region of the spectrum for an initial alignment. Another problem which might occur is that there may be no clearly identifiable single peak. In this case the solution will depend on the particular set of spectra, but will probably necessitate the use of a more complicated peak finding algorithm than that outlined above.

Normalisation

It is necessary to normalise the spectra prior to analysis since intensity differences between different spectra may be caused by instrumental variables such as the fit of the coil rather than differences in absolute values of concentrations. When spectra are acquired *in vitro* it is normal to introduce a reference substance, of which the exact quantity is known, to the sample. The spectra can then be scaled using the intensity of the peak that represents this reference. However, it is not usually feasible to have such a reference when data are acquired *in vivo*. In this case the methods of normalisation are used that will allow for the comparison of differences in the relative proportions of the intensities rather than absolute differences.

Two of the most commonly used standard procedures for normalising vectors are:

- to sum the values in the data vector and then divide each value by this sum
- to sum the squares of the values in the data vector and divide by the square root of this sum.

The first procedure is equivalent to numerically integrating the spectrum and dividing each point by this integral, the second normalises the data vector to unit length. In this work, the second method was used, because the value of the divisor, and therefore the magnitude of the elements of the first method, will be affected by the number of spectral datapoints with negative values. In some spectra, for example those of ^{13}C, some of the spectral datapoints will have negative values due to coupling effects. In principle if there are large numbers of negative values, the integral and thus the divisor could be zero. Indeed if there are a large number of points with a negative value the spectrum can be turned upside down by this procedure!

4.3 Feature Extraction

4.3.1 Requirements

Feature extraction is concerned with finding the best patterns to discriminate and classify the data. In the case of MRS spectra this means choosing appropriate measurements to represent the spectra, and then finding which combination or subset of these measurements provides the best discrimination.

If classification is the main aim, the feature extraction process will involve finding the representation of the data which gives maximum discrimination, and therefore the best classification results for the test set. However, for many applications, including this one, part of the purpose of the feature extraction process will be to identify the biochemical differences between the classes. For this purpose it will be necessary to be able to relate these features back to the original data. It is thus important to try and find features which not only give good discrimination, but are also meaningful. For this reason it may be desirable to pre-process the spectra as little as possible.

In addition to these requirements, the number of features that may be used is limited to the number of samples in the training set due to the dangers of over-fitting that were discussed in Chapter 3. Most authors, for example [Massart *et al.*, 1988] [Kowalski and Wold, 1982], suggest that for linear discriminant analysis the number of variables should be ideally be no more than the number of samples divided by 3.

The feature extraction process is generally the most challenging part of the pattern recognition process, since the methods used will be dependent on the particular type of data. This may mean having to develop methods from scratch for this stage, rather than being able to apply standard methods as we can for the processing and classification stages. While there are standard methods available for processing the magnetic resonance spectra, and classification methods which can be applied to a wide range of types of data, there are only a small number of standard methods available for feature extraction, and some of these are only appropriate for data in which the number of variables is relatively few.

The approach to feature extraction taken in this thesis is to regard the whole spectrum, that is the complete set of spectral datapoints, as the initial set of features or variables, and to successively reduce the number of variables to the optimal number for classifying the spectra.

This is not as difficult as it may first appear, since many of the variables, for example those representing data points in the region of a peak, will be highly correlated with one another. It may be thus possible to select one or two variables from the peak regions or alternatively combine

them in some way. Also many of the variables will be in regions where there are no peaks and will therefore hold no discriminatory information. The feature extraction process will involve finding the best way of discarding the redundant variables and of combining, or selecting a subset of those remaining, in order to best represent features for discrimination.

There are a number of factors that need to be taken into account when considering spectral data. Apart from the fact that datapoints in any peak region will be highly correlated with one another, these datapoints may also be highly correlated with points from other peak regions, either because they represent the same metabolites as do coupled peaks, or because the metabolites themselves are highly related. Many classification methods, including LDA, perform better with uncorrelated variables and this must be borne in mind when choosing features.

Another factor to be considered is that, even after peak alignment, datapoints representing a certain chemical shift are likely to occur at slightly different positions in different spectra. This may be an effect of the digitisation process, or it may be caused by phase differences. It may also be the effect of pH differences (either inter- or extra-cellular) in the tissue being examined. Because *in vivo* spectra have wider peaks (and thus the datapoints representing a certain metabolite are spread over a number of datapoints) the shifts may not affect the values of the variables as much as those acquired *in vitro*. It should be noted however, that although the shifts caused by pH may cause problems in alignment of datapoints, they do provide important information not provided by data acquired *in vitro*.

4.3.2 Strategy

The strategy that was developed for feature extraction was to use a combination of statistical and display methods for preliminary investigation in order to get an idea of the best discriminatory features. Once potential features were identified, the next step was to investigate methods for representing these features as numerical values. In some cases this required extra processing of the spectral datapoints before extraction. The extracted features were then input as variables into the statistical package SPSS [SPSS Inc., 1987], and a correlation matrix was created in order to investigate the relationships between the set of variables. On the basis of this investigation features were either selected without further processing or combined using PCA and entered into the linear discriminant program for the final feature extraction stage.

These steps can be summarised as:

- preliminary investigation of spectra using data display methods to identify potential features
- pre-processing spectra in order to represent the salient features in the spectra
- extraction of features i.e. variables for statistical analysis
- calculation of correlation coefficients in order to investigate relationships between the variables
- calculation of principal components
- selection of best subset or combinations of the features for discrimination.

The following sections describe the methods that were used in this thesis at each of these stages.

4.3.3 Preliminary Investigations

Both data display and statistical methods, or a combination of the two, may be used to find differences between classes of spectra. Methods of data display have the advantage that they give an idea not only of where the differences lie, but also of how these differences might be best represented. If there are large differences between different groups of spectra these may have been identified during the initial investigation when each spectrum was plotted after peak alignment and phasing. This initial inspection is also important for identifying poor quality data, or outliers, that is spectra which are very different from the rest of their group.

Less obvious differences between the classes may be investigated by creating a 'mean spectrum' for each class, and then comparing these by plotting them on the same graph. This can be achieved by calculating the mean values of each datapoint in the spectrum (which had been processed as described above) for each class. The differences, and thus the important regions for discrimination should be discernible on visual inspection of these plots. In order to see how well the groups are separated, the datapoints with the greatest difference in means for different groups can be plotted two at a time on a scatterplot. A lack of any discernible differences between the mean spectra will suggest that it will probably not be possible to discriminate between the groups successfully, since it will indicate that the spectra do not differ significantly between the different groups. However, this will not always be the case. If, for example, the data is bimodal the means of datapoints for different classes which are well separated may possibly be the same. However, there was no reason to suspect that the data investigated for this study had such a distribution. This was confirmed at a later stage in the analyses (see section 4.3.6) by examination of plots of the principal components and other extracted features

Since ideal features for classification will have different means between the classes but small interclass variance it is also useful to calculate and plot the variances for each datapoint for each class.

4.3.4 Extraction of Salient Features from the Spectrum

The decision that must be taken at this stage is how best to represent the spectra in order to select a set of numerical variables for further statistical analysis. This stage will depend very much on the results of the initial investigation. The choice of which features to use for describing and classifying data normally involves a trade-off between the computational feasibility of using low-level data, against the possible added error and information loss incurred in extra pre-processing for higher-level features. MRS data are normally described by measurements of intensities at known resonances, obtained either by measuring the peaks in the Fourier transformed spectrum or by a time-domain fitting method. These measurements are usually obtained by semi-automated means and need considerable processing of the data, as well as subjective judgements by a highly trained operator. In addition the measurements may not always be accurate due to problems such as baseline distortion and overlapping peaks.

Fully automated methods are desirable, not only because they allow for a fully automated system, but also because they remove user-bias, and enable information from the whole spectrum to be utilised. Some of the methods used currently for spectral analysis, in particular some of the time-domain fitting methods, may be, or have the potential to be fully automated. However most current methods are based on measurements of a limited number of frequencies which are chosen on the basis of prior knowledge of the biochemical nature of the data, and therefore do not utilise information from the whole spectrum. In this work I chose to investigate fully automated methods for extracting features from *in vivo* magnetic resonance spectroscopy data which make no prior assumptions about the relative positions and importance of the peaks, but select features purely on

the basis of their power to discriminate between classes.

There are a number of choices at this stage ranging from using the complete set of spectral datapoints to using a few selected peak heights. Approaches used in previous studies, which include, for example, selecting regions of the spectra and either averaging the values of the datapoints in this region or selecting the point with the highest value are reported in Chapter 3 (section 3.5.3). Since feature extraction is one of the key topics of this research, I decided to use a more 'intelligent' approach which involves selecting features purely on the basis of their discriminatory power, only discarding features or datapoints at this stage if it seemed that they were not necessary for discrimination.

Three types of spectral features were investigated in this research. These features were chosen on the basis of the preliminary investigations:

1. peak heights
2. spectral datapoints
3. wavelet coefficients

Peak heights may be a reasonable choice of feature if their mean values differ between classes and they are clearly identifiable. These can be extracted using the average spectrum for the whole group as a template for the peak positions, and then automatically extracting the value of the datapoints at these positions from each individual spectrum. The algorithm is as follows:

1. Plot the means of each datapoint for the training data.
2. Using this mean spectrum as a template to identify the position of the peaks – this can be done either by visual inspection, or automatically using a peak finding algorithm, for example [Abbott, 1994].
3. List the indices of these peaks
4. Using this list, for each spectrum in the data set, select the highest value within two points either side of each indexed point.
5. For each spectrum save these values. This list of values will constitute the feature vector, i.e. the set of variables to be used for further statistical analysis.

The second method is to select a block (or alternatively several blocks) of contiguous datapoints from each spectrum as the feature vector. These points may be either from the whole spectrum, the region of the spectrum which included all the observable peaks, or alternatively datapoints may be selected from certain regions which, on the basis of initial inspection of the individual or mean spectra, appear to hold the most discriminatory information.

The third method requires further processing of the spectrum by carrying out a wavelet transformation on a set (or sets) of contiguous datapoints. In order to carry out this transformation, it is necessary to select 2^n datapoints since the DWT algorithm requires that the length of the data vector is a power of two. The following section discusses the reasons for using this transform and the steps involved in this procedure.

4.3.5 The Discrete Wavelet Transform for Pre-Processing the Spectra

The preliminary investigations of both data sets examined in this study indicated that it may be possible to classify spectra using patterns based on spectral shapes, either of individual peaks or of combinations of peaks. This could be useful when the peaks are difficult to identify and quantify, and could be particularly useful for those spectra with overlapping peaks. It was thus decided to see whether a method based on modelling these shapes might be used to classify these sets of data.

Since no biochemical knowledge was being utilised in this analysis, a method of feature extraction was needed which would make no prior assumptions about the positions of the peaks, but any features would need to be implicitly related to their shape and relative heights. It was thus necessary to find a way of compactly representing these shapes in a succinct way. The discrete wavelet transform provides an excellent way of representing important features, and is also very good for data compression. It thus seemed a good choice for this study. The wavelet transform has been successfully used to extract features for classification of infrared spectra [Bos and Vrielink, 1994] but I believe this is the first reported study of its use for classifying MR spectra.

The wavelet transform can be used to transform a data vector of length n, where n is a power of two, into another vector of length n wavelet coefficients using a set of n orthogonal basis functions called wavelets. Each wavelet coefficient is calculated by taking the dot product of the data vector with one of the basis functions. These basis function are derived from a single function, the 'mother wavelet', by a series of dilations and translations.

As mentioned in Chapter 3, there are an infinite number of potential functions which satisfy the conditions to be a mother wavelet, but in practice only a few are used. A popular class of wavelets, which was used in this thesis, is the Daubechies series [Daubechies, 1992]. A particular set of wavelets will be specified by a set of numbers called wavelet filter coefficients. The most simple (and most localised) member of the Daubechies series, which is often called Daub4, has only four coefficients. By contrast Daub20, which has 20 coefficients is much smoother, but less localised. These difference are demonstrated in Figure 4.1 which shows a single basis function derived from each of the wavelet families Daub4 and Daub20. The plots of these basis functions were created by carrying out an inverse DWT (using Daub4 and Daub20 filter coefficients) of a data vector in which all the coefficients save coefficient number 11 (which had value 1) were set to 0. In this study, I have only used the Daubechies 20 wavelets, and did not attempt to fit the mother wavelet to the problem. This is because I wished to investigate the use of wavelets as a general tool. Both the Daub4 and Daub20 wavelets were investigated at an initial stage. The Daub20 wavelets provided slightly better classification results for this initial study, but the differences did not appear to be critical.

The program used for carrying out a wavelet transform of the spectral datapoints was based on the discrete wavelet transform (DWT) algorithm presented in [Press *et al.*, 1992] which uses the Daubechies wavelet filter coefficients. The algorithm consists of applying a wavelet coefficient matrix (derived from the wavelet filter coefficients) hierarchically, first to the full data vector of length n and then to the "smooth" vector of length $\frac{n}{2}$, then to the "smooth-smooth" vector of length $\frac{n}{4}$ and so on until only two "smooth- $\cdots$ -smooth" components remain. This procedure, which is described more fully in [Press *et al.*, 1992] is a composition of orthogonal linear operations and the whole DWT is itself an orthogonal linear operator.

The resulting vector of wavelet coefficients is divided into components of different smoothness or scale ranges. The first two wavelet coefficients (coefficients 1 and 2) in this vector represent the fully smoothed data vector, which is in effect little more than the mean value of the spectrum. The next two coefficients (3 and 4) represent slightly lower scale features in the data, and the last $\frac{n}{2}$ coefficients represent the very low scale (or high frequency) information in the original data

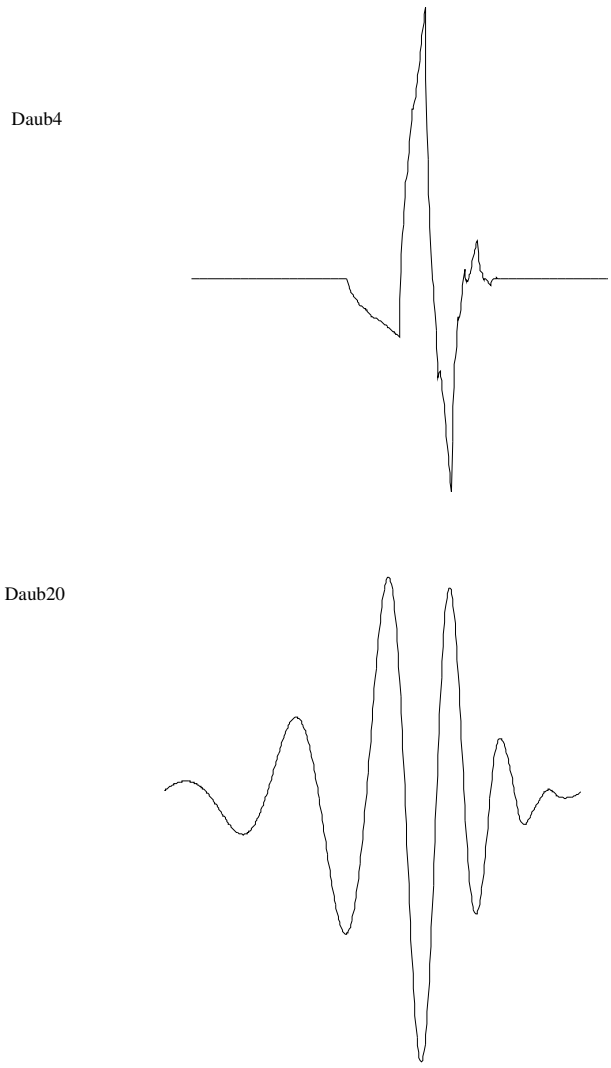

Figure 4.1. Basis functions from the same position and scale range (corresponding to coefficient number 11) from the Daub4 and the Daub20 families. Note how the Daub20 is much smoother, but less localised than the corresponding Daub4 wavelet

vector. For many types of data these coefficients may represent mostly noise in the original data vector and it is often possible to discard these coefficients with little loss of information.

Thus the algorithm presents the wavelet coefficients as a vector filled sequentially with the coefficients of the different scales. Each scale corresponds to a different dilation of the mother wavelet. The lower numbered coefficients represent the large scale features in the spectra and the higher numbered coefficients represent small scale features. For a vector of length 2^j, the coefficients are ordered into j scale levels. Scale level 1 is represented by wavelet coefficients 1 and 2. Scale level i, where $i = 2 \ldots j$, is represented by wavelet coefficients numbered from $2^{i-1} + 1$ to 2^i. Within a scale, each coefficient corresponds to a different translation of the (dilated) mother wavelet. Thus the shape of the most of the wavelets within a scale range will be the same but their position will be different. This means that a coefficient at a certain position within a scale level will represent features at that relative position in the original data vector. This localisation property means that it is possible to relate the wavelet coefficients to features in the original data. For example, features in the centre of the original data vector will be represented by the wavelet coefficients at the centre of each scale level. For example coefficient with index 48 (i.e. $2^6 - 2^5 +$

$\frac{2^6-2^5}{2}$) will represent a central feature at scale level 6. The shapes of the wavelets corresponding to the coefficients at either end of a scale range will have slightly different shapes from the others due to a wrap-around effect.

Transformation of a data vector using code written in C to implement the algorithm in [Press *et al.*, 1992] takes approximately 0.12 seconds of CPU time on a sun SPARCserver 4/690MP. The steps used to carry out a DWT of the spectral datapoints are as follows:

1. Select a set of contiguous datapoints, whose number is a power of two, from the spectrum, to provide a data vector of length 2^n
2. carry out a DWT of this vector to produce a vector of 2^n wavelet coefficients.

The values of this vector were then entered as variables for statistical analysis. Note that this vector is the same size as the original vector of spectral datapoints. However, it is to be hoped that it will be possible to discard many of these wavelet coefficients while retaining the important discriminatory information. This process is described in the next section.

4.3.6 Feature Selection and Reduction

This stage is concerned with reducing the number of features that have been extracted from the spectrum to a number suitable for the discriminant program. Since ideally the maximum number of variables used for the discriminant function should be no more than the number of sample divided by three, this may mean having to reduce the number of variables quite dramatically.

Three methods were used to reduce the number of variables before input to the discriminant program. The first was to select the individual variables that were most highly correlated with the class of the spectrum. The second method was to carry out a PCA of some or all of the variables and then to discard the PC's which accounted for less than a certain percentage of the variance in the data. The third, which was used for the variables representing wavelet coefficients, was to select coefficients from certain scale levels of the wavelet transformed data. For the data sets studied in this thesis a combination of these methods was generally used. For example, scale levels of wavelet coefficients were first selected using the correlation coefficients and then a PCA was applied to the reduced set of wavelet coefficients. Once the number of variables had been reduced sufficiently, a linear discriminant program was used to find the best subset of these variables. Section 4.4 discusses how this procedure was applied.

Feature Selection Using the Correlation Matrix

The object of this stage was to investigate the discriminatory power of individual variables in the feature vector, and to determine which features should be selected, either for direct entry into the discriminant program, or alternatively for further feature reduction using PCA.

If spectral datapoints, or wavelet coefficients, are used as the features it is likely that the number of variables will considerably exceed the number of spectra in the data set. However, some of these variables may represent regions of the spectra where there is either no signal, or no discriminatory information. In addition many of the variables may represent essentially the same information. For example, as the spectral peaks are relatively broad in *in vivo* MR spectra, the number of variables representing a particular metabolite will normally be spread over a number of datapoints. Thus the datapoints from the peak region will represent essentially the same information. This is also the case for datapoints which represent coupled peaks. Thus the intrinsic

dimensionality of the feature space will normally be much lower than the number of features representing the spectrum, and in principle it should be possible to discard many of them with no loss of discriminatory information. The task of the feature selection process is be to filter out such 'redundant' features and to keep only those that will contribute to the classification.

The Problem of Selection Bias Before describing the methods that were used for selecting features, it is worth discussing a potential problem that needs to be taken into account when considering which features to select for the discriminant program. This was the problem of selection bias which was discussed in Chapter 3.

The advantages of using individual variables in the discriminant program is that the features will retain more of their original identity, and can therefore be more easily related back to the original data. The disadvantage of this approach is that such methods may involve selection bias. Selection bias occurs when the subset of variables is not chosen independently of the data used to test the discriminant rule. With data sets which have a large number of samples, this problem can be avoided by splitting the samples into two groups, one of which is used to select the variables and to develop the discriminant rule and the other to test this rule. However, for most studies using medical MRS data the sample sizes are too small to do this and a method such as the leave-one-out procedure must be used for testing the rule. In this case the same set of data will be used both to select the features and to test the discriminant rule. This may lead to an over-optimistic evaluation of the classification rule.

While it is very important to be aware of the possibilities and effects of selection bias, it is not necessarily a major problem. It is often possible to check, for example by consulting with those with an expert knowledge of the biochemistry of the spectra, whether or not the features that have been identified can be related to known biochemical differences. Also, if the highly correlated variables can be related to peak regions in the spectra, it is reasonable to suppose that these correlations do not occur by chance. However, if a highly correlated datapoint appears in a region where there are no peaks, one should be suspicious that this correlation occurs by chance. Although only coefficients which had a significance of $p < 0.01$ were selected, this did not guarantee that these correlations did not occur by chance, particularly when the number of variables was large (e.g. 512), since $p < 0.01$ just means that the probability of a certain correlation occurring by chance is 1 in 100.

Strategy Three methods for feature selection were discussed in Chapter 3. A comparative study [Wu *et al.*, 1995] of these methods showed that similar features were selected by all three methods when they were used for classifying infrared spectra. A preliminary comparison of these methods using the ^{13}C data set analysed in this research confirmed these findings.

In this thesis, features were selected using correlation coefficients. While it appeared that other methods might be equally successful in finding discriminatory variables, the method based on correlation coefficients offers the advantage that it facilitates the investigation of the relationships between all the variables in the data set in addition to those between the spectral features and class. The SPSS program that was used to calculate the correlations is particularly helpful in this respect as it clearly identifies the correlation coefficients with significance levels of $p<0.01$. Thus, even when the correlation matrix is very large it is still possible to investigate the important correlations between variables relatively easily. A limitation of the SPSS correlation procedure is that the number of variables that can be used in the program is 200. Thus in some cases it was necessary to split the vector of spectra features into two or three sub-vectors and repeat the procedure several times.

The strategy used in this thesis was first to assign an appropriate numerical value to each class.

Correlation coefficients can only be used when the variables can be ranked; when there were more than two classes which could not be ranked, each pair of classes was examined in turn. A correlation matrix was created for all the variables and this was examined in order to identify those that were significantly correlated with the class of the spectrum. For many of the pairs of classes examined in this research it was found that a number of variables had high correlations with the class of the spectrum. This was the case for each of the three types of features used to represent the spectra.

Different procedures were adopted according to the type of the feature. These are described in the following three sections.

Selecting Spectral Datapoints The correlation matrix for the variables representing the sets of contiguous spectral datapoints showed that a number of these variables were highly correlated with the class of the spectrum. The variables showing the highest correlations with class were typically in groups of five or six adjacent values. When the positions of these datapoints were related back to the spectrum it was found that they coincided with the positions of the peaks in the spectra, as would be expected. In addition many of the datapoints in these regions were very highly correlated with one another which confirmed that they were from regions representing the same biochemical information (i.e. the peaks in the spectrum). This was also found to be the case for the variables which represented coupled peaks.

Once the variables which were most affected by the class of the spectrum had been identified, it was necessary to choose which of these should be selected for further analysis. This was carried out by first identifying groups of variables (i.e contiguous blocks of variables) which had high correlation coefficients with class number. From each of these regions the variable that was most highly correlated with class was then selected. The value of this variable was then extracted from each feature vector and entered into the discrimination program. For the purpose of selection a region was defined to be a group of at least three variables. Variables were not selected from regions smaller than this due to the possibility that the correlations may have occurred by chance.

After selecting the features and creating the new feature vector, the correlation coefficients between the newly selected variables were examined. If it was found that any pairs of the retained features had correlation coefficients greater than 0.85 in absolute value, one of these features, i.e. the one with the lowest absolute correlation with class, was discarded. This value was chosen because the correlation coefficients of datapoints representing known coupled peaks were observed to have correlation coefficients of at least 0.85.

The steps for selecting the spectral datapoints for further analysis can be summarised as follows:

1. Assign a numerical value for each class and add the class value to each feature vector. N.B this is only appropriate if classes can be ranked - see Section 3.53.
2. Calculate a correlation matrix for all the variables in the feature vector,
3. Identify those variables which have a significant correlation ($p < 0.01$) with the class of the spectrum
4. If more than two contiguous variables have significant correlations with class, and also with each other, select from this group the variable with the highest absolute correlation coefficient
5. Examine the correlations of the selected datapoints with one another. Discard those with correlation coefficients with absolute value higher than 0.85.

Since correlation coefficients can only be used when the variables can be ranked It should be noted that the calculation of correlation coefficients with class which have been arbitrarily assigned a numerical variable is only appropriate if the

Selecting Peak heights For the features representing peak heights many of the redundant datapoints will have already been discarded when selecting the peak intensities. However, if there are a large number of peaks it may still be necessary to reduce these in number for the discriminant program. A similar procedure to that described above for finding the datapoint features which vary most with class was used for selecting the peak height variables. However, in this case it was not necessary to select variables from peak regions since this had already been carried out when extracting the peak intensities.

Selecting Wavelet coefficients As explained in Section 4.3.5, the wavelet coefficients are ordered in scale levels. The first few wavelet coefficients in the data vector will represent the large scale features and the higher indexed coefficients will represent very small scale high frequency features. Preliminary examinations of the correlation matrices for wavelet coefficients indicated that very few of the coefficients from the last two scale levels were significantly correlated with the class of the spectrum. It was therefore decided to discard the coefficients representing the small scale information i.e. the last two scale levels which are represented by wavelet coefficients indexed from $\frac{n}{4}+1 \ldots n$ (e.g. reduce from 512 to 128). It should be noted that discarding all the coefficients from a whole scale level if only a few correlate highly may not always be the best strategy. However, these coefficients were found to give the best results for this study – see Chapter 5.

If the differences between the spectra are subtle, it is possible that some of the higher indexed wavelet coefficients, representing small scale information, may represent features important for discrimination. If the number of samples for data sets examined in this thesis had been larger I would have liked to examine the possibility of using such such small scale features in the discriminant program. However, the potential problem of selection bias precluded such an investigation.

It was found that the wavelet coefficients representing the first few scale levels provided successful classification results for most of the data sets examined in this study.

The steps for selecting the wavelet coefficients can be summarised as follows:

1. Given a vector of wavelet coefficients of length n save the first $\frac{n}{4}$ wavelet coefficients
2. Calculate the correlation coefficients of these wavelet coefficients with the class of the samples.
3. Identify which scale levels contain wavelet coefficients which have correlations of significance greater than $p < 0.01$ with class.
4. Discard all the coefficients from any scale levels which have no, or only a few significant correlations. (N.B. this number will depend on the scale level, since each level will be represented by a different number of wavelet coefficients).
5. The wavelet coefficients not discarded by this process will be the variables of the new feature vector.

Feature Reduction Using Principal Components

The object of this stage was to see if the dimensionality of the feature set could be reduced using PCA. PCA, a very widely used statistical method for feature reduction, transforms the original

variables into a new set of uncorrelated variables called principal components. These new variables are linear combinations of the originals derived in decreasing order of importance so, for example, the first component accounts for as much as possible of the variation in the original data. If the original variables are highly correlated, the first few components will account for most of the variation, and can be used in place of the original variables with little loss of information. It can be used as an alternative to the feature selection methods described above, or it can be used to further reduce the features already selected.

There are a number of advantages of using PCA rather than selecting individual variables. The first is that PCA provides features (the PC's) that are uncorrelated. This is particularly useful for data sets which have a small number of samples, when the number of variables that can be used is very small. This is because a number of uncorrelated variables should, in principle, provide more information than the same number of highly correlated variables. It has the advantage over other methods of feature selection based on class differences in that the components can be selected purely on the basis of the variation they explain again using no class knowledge. This means that no selection bias will be introduced if the test set is used in the feature selection process, as will normally be the case when the leave-one-out method is used.

However, a disadvantage of selecting the PC's on the basis of the variance they explain is that they may not necessarily provide the best features for classification. It is quite common in practice to find that the vector which is most highly correlated with class is one corresponding to one of the smaller PC's [Miller, 1990]. Another disadvantage of this technique is that it is sample dependent and can be unstable if there are a large number of variables compared with samples. Subsequently the inclusion of one or two extra samples may completely change the composition of the PC's.

In this research PCA was only applied directly to the whole feature vector when the number of variables did not exceed the sample size. Otherwise it was applied to a subset of the variables. Although it is possible to use PCA with a larger number of variables, I preferred not to do this due to the instability problem mentioned above. Instead the feature selection methods described above were used to first preselect a subset of the available variables before carrying out a PCA. This proved necessary for the feature vectors of datapoints and wavelet coefficients, where the number of features considerably exceeded the sample size. For these two types of features the number of features was reduced to a number less than the total number of spectra in the data sets and then a PCA was carried out of these variables. In the case of peak heights there were far fewer original variables so it was reasonable to carry out a PCA directly of these.

The correlation matrix was used to calculate the PC's. Once the PC's had been determined, only those which explained a certain percentage of the variance – in most cases 90% – were selected for further analysis, the remainder were discarded. The number of PC's that explained this proportion of variance depended on the particular data sets. However, in all cases this number was much smaller than the number of variables used in the PCA. The steps for carrying out a PCA were as follows:

1. Select a subset of the variables using one of the techniques described above,
2. Carry out a PCA of these variables (using SPSS)
3. Save those PC's which explain 90% of the variance in the data set.

Full details of how PCA was used in this study are given in the next chapter.

Although the main purpose of using PCA in this research was to reduce the number of variables for discrimination, there were two additional reasons why it was useful to carry out this analysis.

Firstly, display of the first two or three principal components for each subject on a scatterplot may be useful for indicating which methods of classification may or may not be successful. For example a compact class next to a disperse one will indicate that LDA is not appropriate but that KNN classification may be [Massart *et al.*, 1988]. The plot should also show if the data is multimodal, which, although unlikely to be the case for this type of data would mean that LDA and also most of the methods that have been described for feature selection will be inappropriate. For the data sets studied here scatterplots of the first few PC's did not indicate that LDA would be inappropriate, however, it provided a useful check.

Another reason for carrying out a PCA is that if the data can be reduced to two or three variables which show good separation of the classes on a scatterplot, it may not be necessary to develop a computer-based classification rule if it is possible to classify the spectra by eye. This plot may be perfectly acceptable to a clinician who wishes to ascertain the relative typicality of a spectrum relative to others in a particular class.

Further Investigations

Once variables have been selected, or reduced, using the methods described above it is useful to investigate how well the classes may be separated on the basis of these variables by plotting the values of two or three of these at a time on a scatterplot. It is also useful to carry out a cluster analysis of selected features in order to see whether there are any obvious clusterings of the data. For example 'rogue' data which have been entered into the analysis by mistake, or alternatively pre-processed incorrectly may be detected using this method. Such an investigation proved very useful in this research for detecting two samples which had erroneously been labelled with the wrong acquisition time.

Other variables such as the sex of the subject, the age, or indeed any other data that might be available, may be added to the feature vector and included in the correlation matrix. This can be very useful for preliminary screening of the data and sometimes surprising factors may be uncovered by examining the correlation coefficients for these variables. Two such factors were discovered in the process of analysing the ^{13}C data in this study. The first was that certain parameters of the MR experiment had been changed half way through the data collection process, without my knowledge! This was discovered when it was observed that the height of one of the peaks was affected, (that is it had a higher correlation than would normally be expected) by the date of data acquisition. Since this could have affected the subsequent analysis it was very useful to discover this fact. The second surprise, arguably less useful, was that one of the small peaks was highly correlated with the sex of the subject and could be used to predict the sex of the subject with an 88% success rate (see next chapter for details).

4.4 Classification and Description

This stage involves developing a classification algorithm to classify the spectra. The two requirements of this stage are

- to develop a rule for assigning a spectrum of an unknown category to a particular class
- to identify the features in the spectrum that provide the best discrimination

Linear discriminant analysis produces a linear function from the variables of known cases which can be used to predict the class of cases whose class membership is unknown. The discriminant functions, like principal components, are a linear combination of the original variables.

However, instead of being calculated to express the maximum amount of variation in the data, they are calculated so as to make the separation between the populations as large as possible. The functions are calculated to minimize differences within a group and maximize them between groups. Chapter 3 (section 3.5.1) gives details of these calculations.

Once the functions have been calculated they can be used to assign unknown individuals to a particular class. Each individual is assigned a discriminant score which is the weighted combination of its values of the discriminating variables. The decision as to whether the particular individual comes from one group or another is based on measuring the distance between its particular score and the centroids (means) of the two different groups, and comparing the probabilities of its membership of each class. This is equivalent to constructing linear decision boundaries between the groups.

A 'training' set of individuals of known class is used to develop the discriminant function and a 'test' set of individuals of unknown class can be used to evaluate how well the functions perform. In this research the 'leave one out' method was used to assess the success of the discriminant functions. This method entails using all the cases (i.e. all subjects), except one, as the training set, and then using the excluded case as the test set. This process is then repeated until each case has been used as the test set.

While there are a large number of methods for classifying data, some of which are discussed in the previous chapter, only linear discriminant analysis was investigated in this research. There were a number of reasons for this. The first was that LDA is a method which has been shown to perform successfully for a wide range of data, including MRS data (see Chapter 3, section 3.5.1). Although this method is optimum when the variables are normally distributed with equal covariance matrices it has been shown to be relatively robust to departures from these assumptions. The second reason was that the sample-to-variable ratio of the test data meant that most of the other methods would be impracticable. For example, nonparametric methods such as nearest neighbour classification, or nonparametric discriminant analysis rely on densely populated feature spaces and thus need a reasonably large number of samples. Another reason is that it is generally preferable to try the simplest approach first, only going on to more sophisticated methods if this does not work. Since was possible to classify successfully both sets of data studied in this research using LDA, there was no need to try more complicated methods for the purpose of this study, of which the foremost aim was to investigate methods of automated feature extraction. A potential problem with the more complicated techniques is that the more parameters that need to be estimated the more the analysis will depend on the training data and also possibly the subjective judgements of the operator.

Another reason for choosing LDA was that it is very quick to run and therefore provides a good methods for investigating different sets of features. For example carrying out a discriminant analysis of 6 variables for 75 cases for three classes of spectra, using the leave-one-out method for the test set (which involved running the program 75 times) took approximately 40 seconds of CPU time.

LDA can be used both to develop the classification rule and to select the best subset of the available variables for discrimination. Most statistical software packages, including SPSS, include methods for subset selection included in the discriminant program. However, this option was not used in this research as it was reasonably straightforward to find the best subset of variables by trial and error, using the correlation coefficients as a guide to which variables would provide the best combinations for discrimination.

The discriminant analysis program first determines the linear decision boundaries and the decision rule using the procedure described in Chapter 3. The boundaries and rules are developed

using the training set of cases whose class is known. A score is then assigned to each spectrum of unknown class and this spectrum is assigned to a particular class on the basis of this decision rule. Then the probability of that spectrum belonging to each class is estimated using equation 3.2. Details of the SPSS program DISCRIM are given in the manual [Norusis, 1994]. This program provides:

- the coefficients of the discriminant functions
- the classification results for each case (subject) together with the estimated probabilities of the case belonging to each class
- the correlation coefficients of the linear discriminant function with each variable included in the analysis
- the percentage of correctly classified cases for both the test and the training sets.

The steps for carrying out LDA were as follows:

1. Select a subset of the variables (on the basis of their correlation with the class of the subject) and carry out a LDA of the whole data set,
2. Note the classification results, and also which of the variables which are most highly correlated with the discriminant function (or functions when there are more than two classes)
3. Drop the variable which is least correlated with the discriminant function and repeat the analysis.
4. If the results are improved or unchanged, add another variable and repeat
5. If not, replace that variable and remove the next least correlated variable
6. Continue this process until the classification results show no change and all the variables have been added
7. Reduce the number of variables to approximately the number of spectra in the smallest class divided by 3 by dropping those which have the lowest correlation with the discriminant function.

This part was used to develop the discriminant rule. The following steps were used to test this rule:

1. Using the selected variables carry out LDA on all but one of the cases, using the remaining case as the test set
2. Repeat until each case has been tested.

4.5 Implementation

All the procedures for transforming and processing the spectra were implemented in the programming language C++. This is an object-oriented programming (OOP) language which facilitates modular software design by allowing the programmer to structure the programs in terms of classes

which define 'objects'. Each class acts as a template for an object, and all the procedures (methods) and variables (called instance variables) that are needed to manipulate the object are 'contained' within the class. Thus each class is composed of a constructor which enables the object to be created, a list of instance variables and a set of methods for manipulating these variables. In addition OOP provides a mechanism called 'inheritance'. This allows 'child' classes to be designed which inherit all the methods, and variables of the parent class, and to which new specialised methods can be added.

The main class of the program was the class 'Spectrum'. The variables for this class included an array to store the datapoint objects, together with variables such as the number of datapoints, reference frequency of the spectrum and index of highest peak. Methods were defined for each data manipulation procedure, and included those for phase adjustment, normalising the spectrum and aligning the peaks. Methods were also included for carrying out the DWT and extracting the values for statistical analysis, and outputting these into a file in a suitable format for SPSS. Using the inheritance mechanism two specialised classes ratSpectrum and CarbonSpectrum, defined extra variables and methods that were needed for the two different types of spectrum, for example for the extraction of the different types of variables for statistical analysis.

In addition to the spectrum classes, two other classes were defined. The class FID defined procedures for Fourier transforming and processing the FID's. The structure of the FID class was similar to the Spectrum class. Each object of the class being defined by an array of datapoints. The datapoint objects for both the Spectrum and FID classes were defined by the class DataPoint. The main variables for this class were the real and imaginary values of each point. The methods of the class provided procedures for manipulating these values.

The code for implementing the DWT and Fourier transform (the fast Fourier transform) was based on the algorithms described in [Press *et al.*, 1992]. The code for phase-adjustment was an implementation of the algorithm described in [Chen and Kan, 1988]. Software (written in C) for extracting the FID's from the ^{13}C and ^{31}P datafiles were provided by Hammersmith and St George's Hospitals respectively. This code was adapted to provide methods for constructing the FID and Spectrum objects used in the program.

The statistical package SPSS was used for the statistical analysis, i.e. PCA, calculation of correlation coefficients and LDA. The programs CORR and NONPAR CORR were used to calculate Pearson and Spearman coefficients respectively. The program FACTOR was used to calculate the principal components. The program DISCRIM was used for LDA. A specialised macro was written, using the macro facility of the SPSS language, in order to implement the leave-one-out procedure for testing the LDA program.

Plots of the spectra were created using the Xgraph program and scatterplots were produced using the Xgobi package. Xgobi is a particularly useful program which facilitates the investigation of relationships between variables by allowing three dimensional plots which can be rotated. It also provides a method for displaying one-dimensional data in the form of a 'dot plot'. This is a device which randomly spreads the points out along a second axis for display purposes only. Most of the procedures used to develop the system could be carried out reasonably quickly. Once the system had been developed it would take only a few seconds to process a single spectrum, carry out a wavelet transform and assign a classification to that spectrum.

4.6 Summary

This chapter describes the development of a classification system for magnetic resonance spectra in which all the processing, i.e. filtering, phasing, peak alignment feature extraction and classifi-

cation, is fully automated. The methods that were used are discussed in detail, together with the issues and problems that needed be dealt with at each stage of the development.

The main aim was to investigate methods that can be used to reliably and automatically classify MRS data, using features that are extracted automatically using the whole spectrum, rather than the selected metabolite resonances which are normally used to describe MRS data. Methods are suggested for selection of features purely on the basis of their power to discriminate between different types of spectra, using no prior knowledge of biochemistry.

Three types of features are suggested: peak heights, spectral datapoints and wavelet coefficients. Which type of feature is appropriate will depend on the particular set of data. Once the features have been chosen and extracted from the spectra, correlation coefficients can be used to select which features to use for the discriminant program, or for further feature reduction using PCA.

Because pre-processing of the spectra and extraction of discriminatory features require methods specific to the particular type of data, the chapter concentrates most on these two stages. This necessitated finding features which could represent the discriminatory information and which can be extracted automatically.

A summary of the steps that need to be carried out at each of the stages in the development is given below.

Stage 1 Spectral Processing

1. Pre-process FID by zero filling line broadening and removing dc component
2. Fourier transform FID to produce a Spectrum
3. Phase adjust
4. Inspect each spectrum
5. Align the peaks and reduce spectrum to region containing peaks
6. Normalise each spectrum

Stage 2 Feature Extraction

1. Plot mean spectra and identify differences
2. Carry out wavelet transform (optional)
3. Extract variables from spectra
4. Investigate correlations between variables
5. Calculate principal components of selected variables (optional)
6. Select subsets of variables
7. Select a subset of the extracted features using LDA

Stage 3 Classification

1. Carry out a LDA using these values and test the resulting discriminant rule using the leave-one-out procedure.

The system was developed and tested using two sets of data, and therefore the methods that were used, particularly those for feature extraction, were influenced by the nature of these data sets. However, because the aim of this research was not to classify one particular set of data but to investigate general methods which can be used to discriminate between different types and classes of *in vivo* spectra, very little reference is made to the particular data sets in this chapter. Instead this discussion is left to the next chapter where the results of applying these methods to the two sets of data is described in detail.

Chapter 5

Results

The previous chapter discussed the development of a prototype system to classify MRS data. Its main purpose was to discuss in detail the practical problems that are involved in designing a system for automatically classifying MRS data, together with the methods that can be used to for solving these problems. The aim of this chapter is to describe in detail how the methodology described in the previous two chapters was applied to the two sets of *in vivo* data:

- a set of 75 ^{13}C spectra obtained *in vivo* from healthy human volunteers of adipose tissue in the leg
- a set of 55 ^{31}P spectra obtained *in vivo* from tumorous and normal tissues in rats.

The two specific aims of the work described in this chapter were:

1. to see whether it was possible to design a fully automated system for classifying these spectra and
2. to develop and test methods for automatically extracting features which could be used to discriminate between the different classes of spectra.

The initial investigation and development of the system was carried out using the ^{13}C data set. This data was very suitable for a preliminary study as the signal-to-noise ratio was relatively high and because it had already been ascertained that it was possible to discriminate between the the two main groups reasonably well by visually inspecting the spectra.

The second set was more challenging, in that the signal-to-noise ratio was much lower, and also because it was not easy to discriminate between the different groups by visual inspection of the individual spectra. None the less good results were obtained for both sets of data. These results have been previously reported in [Tate *et al.*, 1995] [Tate *et al.*, 1996a] [Tate *et al.*, 1996b].

5.1 The Diet Study

This data set was acquired at Hammersmith Hospital as part of a study to examine how the types of fat stored in the body are affected by diet. 75 spectra were analysed. The volunteers were all

normal and healthy and were classified as being either vegan (class 1, n=33), vegetarian (class 2, n=8) or omnivore (class 3, n=34), according to their stated dietary group. Vegetarians were categorised as those who ate no animal flesh (including no fish), and vegans as those who ate no animal products, for example eggs, cheese or milk. [Thomas *et al.*, 1995].

The aim of the analysis carried out in this thesis was the development of a system to classify the spectra according to the dietary group of the subject.

5.1.1 Data

The data consisted of a set of unlocalised coupled ^{13}C spectra obtained on a Picker prototype MRS system operating at a field of 1.5T using an 8 cm surface coil positioned on the human thigh. The coupled spectra were acquired using a 90° pulse at a TR of 30 seconds such that the peaks were fully relaxed. There were 512 points with a dwell time of 100μs per point. This set of data was obtained at Hammersmith Hospital and had already been Fourier transformed when it was received. No apodisation, line broadening or resolution enhancing filters had been employed prior to Fourier transformation. The spectra had been zero filled to 4096 points.

The analysis for this set of spectra was carried out in two stages since initially only half the data (set 1) was available as the other half (set 2) had not yet been acquired. The first set of data was used for all the initial investigations and also for selecting which peak height features to use in the discriminant program. Since the classification results were very similar for both data sets this section describes the results of the analysis for the whole data set.

5.1.2 Spectral Processing

Since this set of data had already been Fourier transformed and zero filled, and because no apodisation appeared to be necessary, the first procedure to be carried out was to adjust the phase of the spectra. All spectra were adjusted automatically for zero order phase prior to pattern recognition analysis using the algorithm described in [Chen and Kan, 1988].

Further processing was carried out to make the spectra compatible for PR analysis, using the methods that were described in Chapter 4. Initial inspection to the spectra showed that each spectrum had one peak that was considerably larger than the others and this was selected for the peak alignment algorithm. This was achieved using the algorithm described in Chapter 4, by finding the index of the highest point in each spectrum and then retaining 511 points upfield of this peak and 512 points downfield thus reducing the number of points to 1024. There was no need in this case to find the spectral region for this peak since it was always the highest point in the spectrum (shown as peak 18 in Figure 5.1 below).

The resulting vector was normalised to unit length to compensate for arbitrary (vertical) scaling differences. A visual inspection was carried out to check that the same peaks were aligned together. This was found to be the case for each spectrum in the study (to within approximately 0.6 ppm), since the largest peak always occurred in the same position relative to the other peaks.

5.1.3 Extraction of Features from the Spectra

Initial Investigation

In order to get some idea of how the spectra from the three classes differed, and to help identify the discriminatory features, a mean spectrum was created for each of the three classes: vegan,

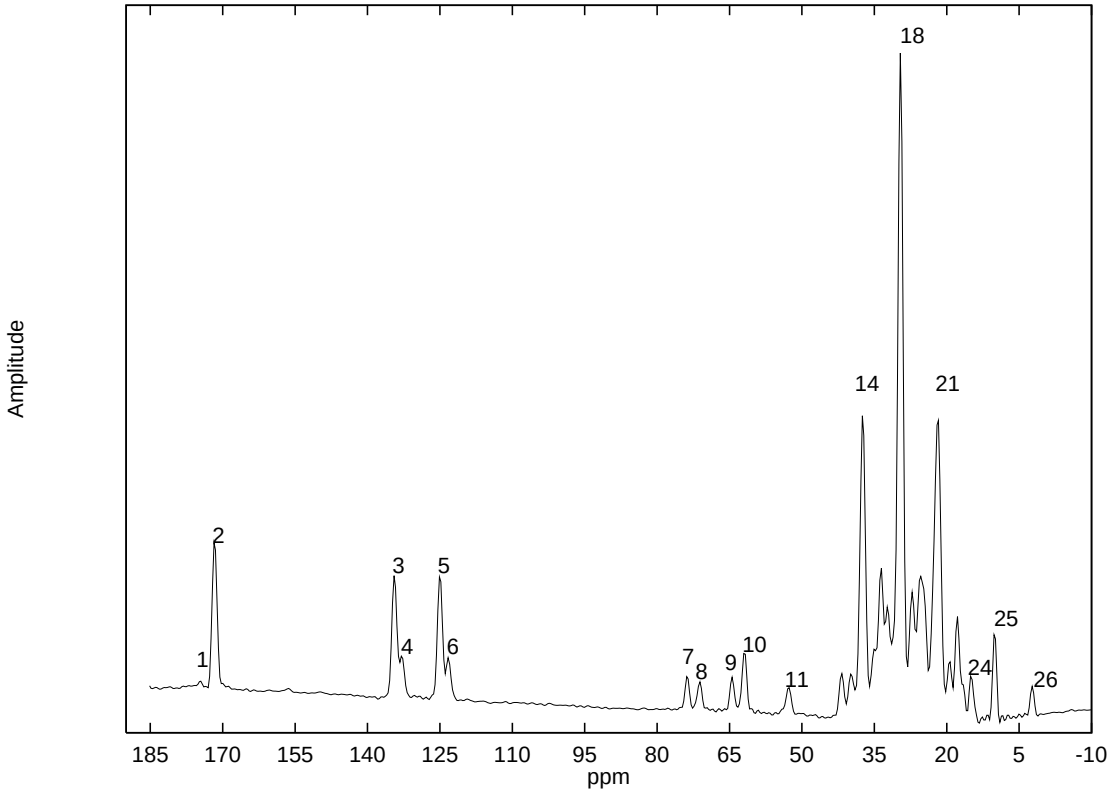

Figure 5.1. Mean ^{13}C spectrum of adipose tissue identifying the 26 peak variables.

vegetarian and omnivore. This was achieved by calculating for each class the mean values of each datapoint in the spectra (which had been processed as described above). A mean spectrum was also created for the complete set of 75 spectra showing that 26 peaks could be clearly identified. These peaks were labelled 1 to 26. Figure 5.1 shows these peaks on the mean spectrum.

The mean spectra for each of the three classes (which all showed these 26 peaks) were then examined by plotting them on the same graph. This plot showed that significant differences could be observed by eye between each of the groups; the main variations between the groups could be observed in the intensities of peaks 3, 4, 5 and 6. Figure 5.2 plots these mean spectra in the region that includes these four peaks, showing that the means of the intensities of these peaks are higher for the vegans than the other two groups.

Lesser differences could be observed in the other peaks and for most peaks the mean spectrum for vegetarians lay between those of the other two extreme classes. The results of this initial inspection suggested that the groups could be distinguished on the basis of peak heights. It was thus decided to extract the values of each of these 26 peak heights from each spectrum in order to see whether they might provide features for classification.

Extraction of Peak Heights

The peak heights were automatically extracted from each individual spectrum of the study using the average spectrum for the whole group as a template for the peak positions. Twenty-six peaks were identified and labelled `p1` to `p26` corresponding to the peak numbers in the labelled spectrum in Figure 5.1. The index of the datapoint representing each peak, i.e. the highest value in each peak region, was first identified and then the 26 values were automatically extracted from each spectrum using these indices. The identification of the 26 peaks was initially carried out by visual

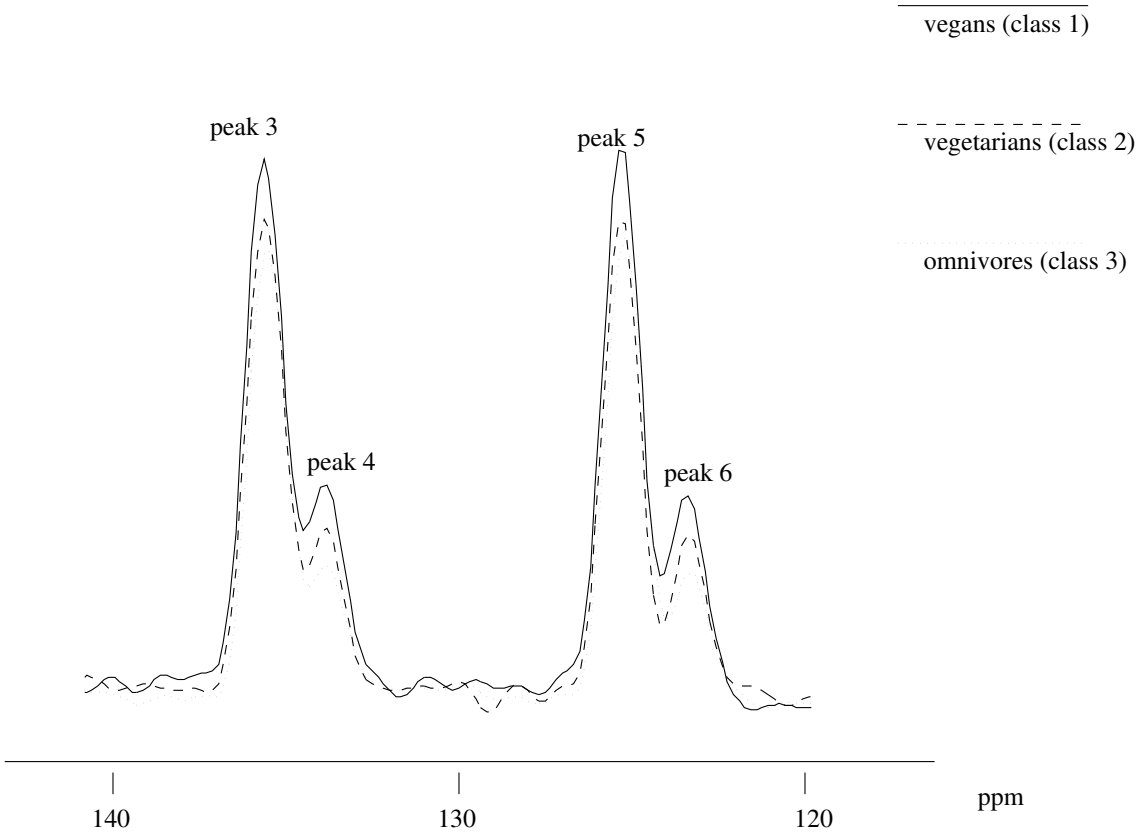

Figure 5.2. Overlay plot of the mean spectra for the three classes (vegan, vegetarian and omnivore) showing the region that includes peaks 3–6.

inspection, but the process was subsequently automated using a peak finding algorithm [Abbott, 1994]. To account for slight differences in the positions of the peaks the program selected the maximum value within two points on either side of the indexed point. Once these values were extracted a scatterplot was produced of the variables `p5` and `p6` for each individual. This scatterplot (Figure 5.3) indicates that a good separation between the two main groups might be obtained the using either or both of these variables.

Pre-Processing the Spectra Using the Discrete Wavelet Transform

The initial investigation also suggested that it might be possible to classify the spectra using patterns based on spectral shapes, either of individual peaks or of combinations of peaks. It was thus decided to see whether the DWT might be used as a method of processing the spectra in order to represent these shapes succinctly.

In order to see whether wavelets might be successfully used as features for classification, 512 points were extracted from each (scaled and normalized) spectrum by selecting 1024 points from the region -10 to 140 ppm, and taking every other point (note that this did not result in a loss of information since this zero-filling had been applied to the FID's). These 512 datapoints were then transformed into a set of 512 wavelet coefficients, numbered from 1 to 512 by convolving the basis functions with the spectrum using the wavelet transform algorithm described in [Press *et al.*, 1992]. As discussed in Chapter 3, this algorithm presents the wavelet coefficients as a vector filled sequentially with the coefficients of the different scales. For this vector of length 512 (i.e. 2^9) the coefficients are ordered into 9 scale levels – each scale level representing the data vector at a certain resolution. Scale level 1 is represented by wavelet coefficients 1 and 2. Scale level i,

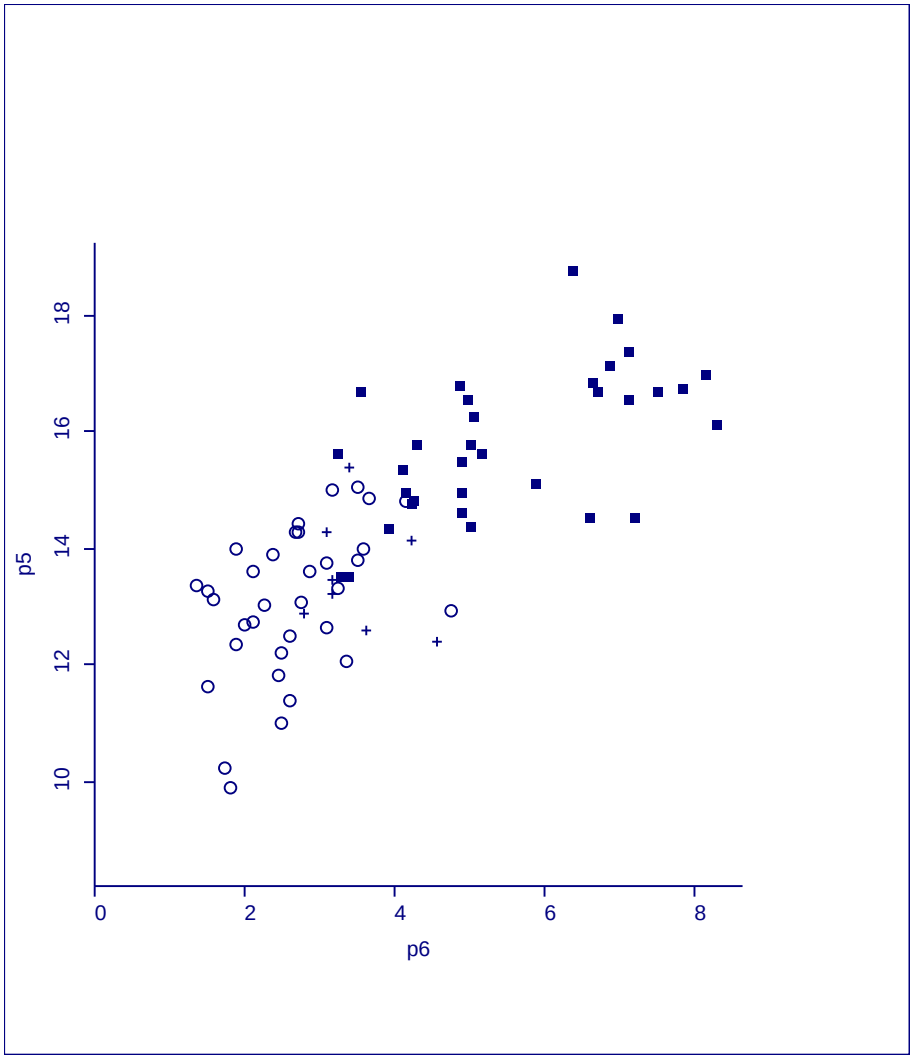

Figure 5.3. Scatterplot showing the value of `p5` and `p6` for each individual. □ represent the vegans, + the vegetarians and ∘ the omnivores.

where $i = 2 \ldots 9$, is represented by wavelet coefficients numbered from $2^{i-1} + 1$ to 2^i.

Figure 5.4 shows a typical spectrum and its wavelet transform. Coefficient numbers are shown in powers of two to mark the different scale levels. This figure shows that wavelet coefficients 256-512 have very small values compared with the higher indexed coefficients. It also shows that the coefficient with the largest intensities tend to be positioned towards the right-hand side of each scale range. This is because the region with the highest energy in the spectrum, i.e peaks 12–26, appear at the right-hand side of each spectrum.

5.1.4 Feature Selection and Reduction

The peak variables (1–26) and wavelet coefficients (1–512) together with the class (1, 2 or 3), sex (labelled 1 for male and 2 for female) and data set number (1 or 2) of each individual were entered into the SPSS package which was used for all subsequent statistical analyses. [SPSS Inc., 1987]. The following sections describe the methods that were used to reduce the number of variables for the discriminant program for the two types of features: peak heights and wavelet coefficients.

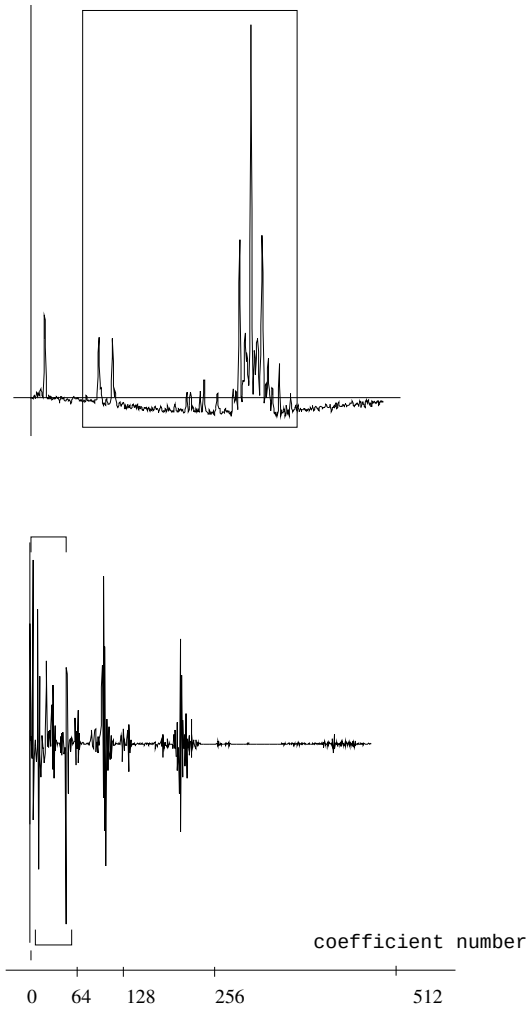

Figure 5.4. A typical spectrum and its wavelet transform; the boxed area of the spectrum shows the region that was transformed, and the marked wavelet coefficients (numbers 3 to 64) indicate those used for classification.

Peak heights

Correlation coefficients were calculated for all the variables, including the sex, class and set number of the spectra. The correlation matrix showed that the peak variables `p3` to `p6` were most highly correlated with subject class. These correlation coefficients, which had values of -0.65, -0.75, -0.72 and -0.76 respectively, were all negative indicating that the heights of the peaks 3 to 6 decrease with class. This confirmed the results of the preliminary observations of the mean spectra which showed that the intensities of the peaks 3–5 were higher for the vegans (class 1) than the vegetarians (class 2), which in turn were higher than those of the omnivores (class 3).

Some of the other peak variables were also significantly correlated with class, but the absolute correlations were all below 0.5, except for those of `p18` and `p20` which were 0.53 and -0.53 respectively. Most of the correlations were negative, with the exceptions of `p14, p18, p21` and `p24 to p26`. Some of the pairs of peak variables were very highly correlated with one another (some with absolute values as high as 0.9). These included `p3` with `p5` and `p4` with `p6` (Figure 5.1). Some of the other peak variables were also highly correlated with one another, which indicated that it might be possible to reduce the number of variables representing the peaks considerably using PCA.

Pearson correlation coefficients were used for this analysis since these were more appropriate for all the variables except class, and it had been ascertained that the Spearman and Pearson correlation coefficients for the peak variables with class had very similar values.

It was interesting that the very small peak (p1) at the low field of the spectrum had a significant ($p<0.01$) correlation of -0.64 with sex, i.e. the height of peak 1 is higher for males than for females. The origin of this peak is unclear, but appears to be related to muscle. Thus subjects with relatively low fat content generally showed relatively higher intensity for this peak.

Another interesting result was that (p1) had a significant correlation (0.7) with the data set number. I subsequently discovered that this was due to a slight change in the acquisition time of the spectra. Fortunately this did not appear to affect the intensities of any of the other peaks, but this peak was subsequently dropped from the analysis. This peak was also not included in the region chosen for the wavelet analysis.

On the basis of these correlation coefficients p5 and p6 were selected as individual variables for the discriminant program. p3 and p4 were excluded, because of their high correlations (> 0.85) with p5 and p6 respectively.

Since many of the variables were highly correlated, this data set seemed a good candidate for PCA. Principal components were calculated using all the 25 peak variables. In order to see whether the spectra could be classified using the more complex region of the spectra which did not include p5 and p6 PC's were also calculated using p14 to p26. The principal components which accounted for 90% of the variance (the first 9 for the whole spectrum and the first 5 for p14 to p26 were then selected for entry into the discriminant program.

5.1.5 Classification Results

The main purpose of the original study of this data set had been to investigate differences between the two extreme dietary groups, omnivore and vegan. Because of this, and also because of their small number, the vegetarians were initially excluded from discriminant analysis. Using the leave-one-out method for obtaining the training set and the procedure described in Chapter 4 for finding the best discrimination variables, 93% of the unknown cases were classified correctly (5 wrong out of 67). The best discriminating variables were found to be p5 and p6; the best discriminant functions being a linear combination of these two variables. P3 and p4 could equally well have been used as the discriminating variables, since they represent coupled peaks.

The discriminant analysis was then repeated using the principal components for the 25 peak variables. When the first five components (which accounted for 80% of the variance) were used as the variables in the discriminant program, 89% of the cases were correctly classified (7 wrong). The same results were achieved using the 5 principal components which had been calculated using just the second half of the spectrum, which showed that the spectra could be classified quite successfully using the peaks less highly correlated with class.

When discriminant analysis was carried out on all three groups, there were a larger number of misclassifications. This may have been due to the fact that the differences in diet between the vegetarians and the other two groups is not so clear-cut, and but also may be due to the fact that the number of vegetarians was much smaller. The results from linear discriminant analysis using peak heights as the variables are summarized in Tables 5.1 and 5.2.

Using LDA it was also possible to classify the spectra according to the sex of the individual. When p1 alone was used as the discriminating variable 88% of the cases were correctly classified.

Table 5.1. Classification results using peak heights as the variables in the discriminant analysis program.

Actual Class	Total Cases	Percent of Cases Assigned to Each Class	
		Vegan	Omnivore
Vegan	33	*87.9% (29)*	12.1% (4)
Omnivore	34	2.9% (1)	*97.1% (33)*

Percentage of groups correctly classified: 93%

Table 5.2. Classification results using peak heights when vegetarians were included.

Actual Class	Total Cases	Percent of Cases Assigned to Each Class		
		Vegan	Vegetarian	Omnivore
Vegan	33	*75.8% (25)*	18.2% (6)	6.1% (2)
Vegetarian	8	12.5% (1)	*75.0% (6)*	12.5% (1)
Omnivore	34	0.0% (0)	29.4% (10)	*70.6% (24)*

Percentage of groups correctly classified: 73%

Wavelet coefficients

The correlation matrix showed significant correlations ($p<0.01$) between some of the wavelet coefficients and class. The greatest correlations were shown by coefficients 35 (-0.67), 37 (-0.67), 38 (0.73) and 39 (-0.71). These coefficients, because they occur at the beginning of the scale level indexed by $33-64$, represent the low field region of the transformed spectrum containing peaks $3-6$ as would be expected. Significant correlations were also found for the wavelet coefficients related to the more complex region of the spectrum (e.g. coefficient 59 (-0.63) near the end of this scale level). This is an encouraging result since it corresponds to an area of the spectrum where the peaks are not easily resolved. The most highly correlated coefficients, i.e. those with correlations greater than 0.5 in absolute value, were nearly all in the range 3 to 64, indicating that it might be possible to use a combination of these 62 wavelet coefficients to discriminate between the spectra.

Many of these 62 wavelet coefficients were also highly inter-correlated, both within and between the scale levels. It was therefore decided to apply PCA to these coefficients in an attempt to further reduce the dimensionality. Although the first two principal components accounted for only 40% of the variance in the data, the second principal component was highly correlated with class (0.8). Analysis of this second principal component revealed that it was significantly correlated with the wavelet coefficients associated with the low field region of the spectrum. Figure 5.5, which shows the values of this principal component for each case shows that very good separation can be obtained between the two main classes on the basis of this one principal component.

The first 18 principal components which accounted for 90% of the variance in the data were input as the variables to the discriminant program. 4 out of the 67 cases from the two main groups were misclassified, a success rate of 94%. When the vegetarians were included, 15 out of the 75 cases were misclassified which represented a success rate of 80%. The results are summarized in Tables 5.3 and 5.4 and demonstrate an improvement over those when peak heights had been used as the variables.

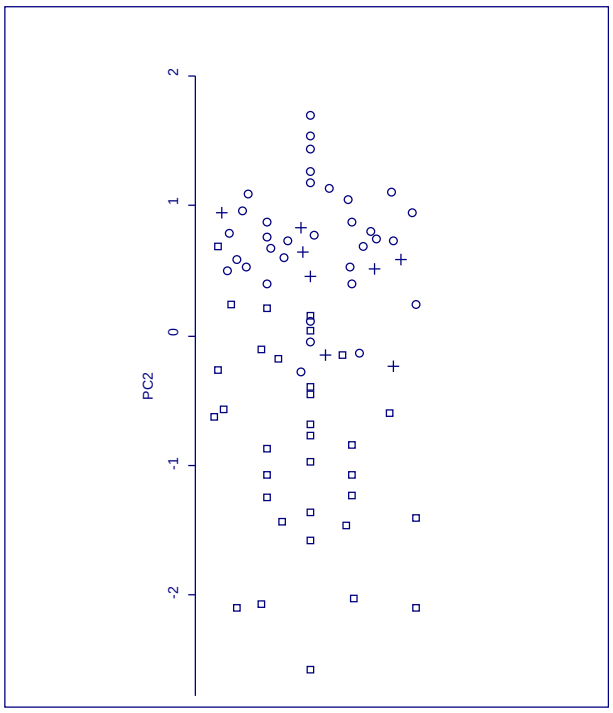

Figure 5.5. Dotplot showing principal component 2 of the wavelet coefficients. This component was highly correlated with dietary class. □ represents the vegans, + the vegetarians and ∘ the omnivores.

Table 5.3. Classification results using linear discriminant analysis employing wavelet coefficients.

Actual Class	Total Cases	Percent of Cases Assigned to Each Class	
		Vegan	Omnivore
Vegan	33	*90.9% (30)*	9.1% (3)
Omnivore	34	2.9% (1)	*97.1% (33)*

Percentage of groups correctly classified: 94.0%

It should be noted that at a preliminary stage of the study the wavelet coefficients numbered from 65 –128 were included in the principal component analysis and discriminant analysis procedure described above. However, the results obtained using coefficients 3– 64 gave better results which justified excluding the higher numbered wavelet coefficients from the subsequent analysis.

5.1.6 Discussion

The results obtained from this set of data show that using linear discriminant analysis it was possible to separate the two main dietary groups successfully, either using peak heights or wavelet coefficients. It was possible to do this using a fully automated procedure for all the processing and analysis of the spectra.

The results using peak heights showed that it was possible to classify successfully this set of spectra using either two selected datapoints or the first five principal components of all 26 peak values. The spectra could also be classified using only peaks 12–26, a fact which indicates that

Table 5.4. Classification results using linear discriminant analysis employing wavelet coefficients when the vegetarians were included.

Actual Class	Total Cases	Percent of Cases Assigned to Each Class		
		Vegan	Vegetarian	Omnivore
Vegan	33	*84.8% (28)*	9.1% (3)	6.1% (2)
Vegetarian	8	25.0% (2)	*50.0% (4)*	25.0% (2)
Omnivore	34	2.9% (1)	14.7% (5)	*82.4% (28)*

Percentage of groups correctly classified: 80.0%

these methods may be useful for analysing more complex spectra. The expert analyst can usually determine the dietary group of the subject by standard methods on the basis of peaks `p3` to `p6`. However, this is not normally possible when these peaks are excluded.

The results using the wavelet transform showed that the spectra could be classified completely automatically with no need for the locating any peaks. These results suggest that the wavelet transform might prove a useful tool for classifying groups of spectra that can be discriminated on the basis of line-shape or spectral pattern. It may also prove useful for classifying spectra with rolling baselines. A feature of the DWT is that the mean level is represented by the first two wavelet coefficients. While these coefficients contain large scale information about the spectra, they may not be necessary for classification, as was indicated in this study. These results also showed that PCA could be used to reduce the number of variables for discrimination successfully, especially when it was used in combination with the wavelet transform. These two techniques can be used to achieve the same objective but the fact that they do so in completely different ways means that they can be used in combination very effectively. An advantage of using both of these methods is that the coefficients can be chosen without using knowledge of the class of the spectra thus removing the possibility of selection bias.

The next section discusses how similar techniques were applied to a set of ^{31}P spectra. These spectra provided more of a challenge than the ^{13}C data since the SNR was much lower, and also because the peaks were less clearly defined and less easy to identify. Another problem was that the number of samples was much small – on average there were only 10 samples in each class. Nonetheless good results were obtained for this set of data when the classes were examined in pairs.

5.2 Study of ^{31}P Spectra from Normal and Cancerous Tissues in Rats

5.3 Data

Three tumour types were studied: Morris hepatoma 7777 (fast-growing, poorly differentiated) and Morris hepatoma 9618a (slow-growing, well differentiated) were grown in female Buffalo rats; Walker 256 carcinosarcomas were grown in female Wistar rats; and GH3 prolactinomas ware grown in female Wistar-Furth rats. In all cases, tumours were implanted subcutaneously into the flank, and generally grew to a size suitable for spectroscopy within 2–3 weeks. The animals used for studies on normal tissues were male Wistar rats.

Spectra from two classes of normal tissue (10 livers and 10 brains) and four tumours (10 h9618a hepatomas, 13 Walker carcinomas, 4 h7777 hepatomas and 8 GH3 prolactinomas) were

obtained from anaesthetised rats (i.p. Sagatal) in a SISCO 4.7T spectrometer using ISIS localisation (volumes 0.22–1.0cm^3, chosen after examination of ^{1}H images) with a 25mm surface coil.

5.3.1 Spectral Processing

The FID's were processed by removing the dc offset (calculated as described in Chapter 4) and filtered using an initial line-broadening factor of 30 Hz (which was later reduced to 10 kHz – see section 5.3.3). The number of points, which was initially 4096 was reduced to 1024 since the apodisation had the effect of cancelling out any signal that might have been present in this region.

After Fourier transformation each spectrum was automatically phased and then normalised by summing the squares of all the values in the spectrum and dividing each value by the square root of this sum. The peaks were then aligned using the highest datapoint (i.e. peak) in the region indexed from 760 to 860 in the data vector The new index for this peak was 461 and the number of datapoints was simultaneously reduced to 512, which covered the region containing all the main peaks. This peak was chosen for alignment because a) this is known to be one of the peaks least affected by pH shift, and b) the initial investigation of the spectra revealed this to be the most clearly identifiable peak in the study.

5.3.2 Extraction of Features from the Spectra

Initial Investigation

A mean spectrum was created for each class. As with the ^{13}C spectra there appeared to be a number of clearly identifiable peaks. Figure 5.6 shows the mean spectrum for the largest group, the Walker's carcinomas. However, although there were a number of clearly identifiable peaks the positions of these peaks appeared to shift in position for the different groups. Also there appeared to be a number of less clearly identifiable peaks on the sides of some of the main peaks. For this reason, it did not seem to be appropriate to use peak intensities to represent the spectra. Instead, a different approach was taken and the 512 datapoints were entered into the SPSS package for further statistical analysis.

The initial stage of the statistical analysis of the spectra was carried out using these 512 values. The second stage of this analysis involved an investigation of the region of the spectrum which is known to be affected by the metabolic changes that occur in cancer, the PME region. This region which contains the lipid metabolite peaks (peaks A–C in Figure 5.7 shown below) was of particular interest to the biochemists who acquired the data. The analysis of this region involved further processing the data by extracting the 32 datapoints indexed from ppm 7.6–5.73 from each spectrum, normalising these 32 points as above, and transforming them using a wavelet transform.

The 512 datapoints and the 32 wavelet coefficients, together with the class number for each tissue type (numbered from 1 to 6) were then entered into the SPSS program for further statistical analysis.

5.3.3 Feature Selection and Reduction

Datapoints

Each pair of tissue types was analysed in turn, and correlation coefficients were calculated between tissue type and each of the 512 spectrum values.

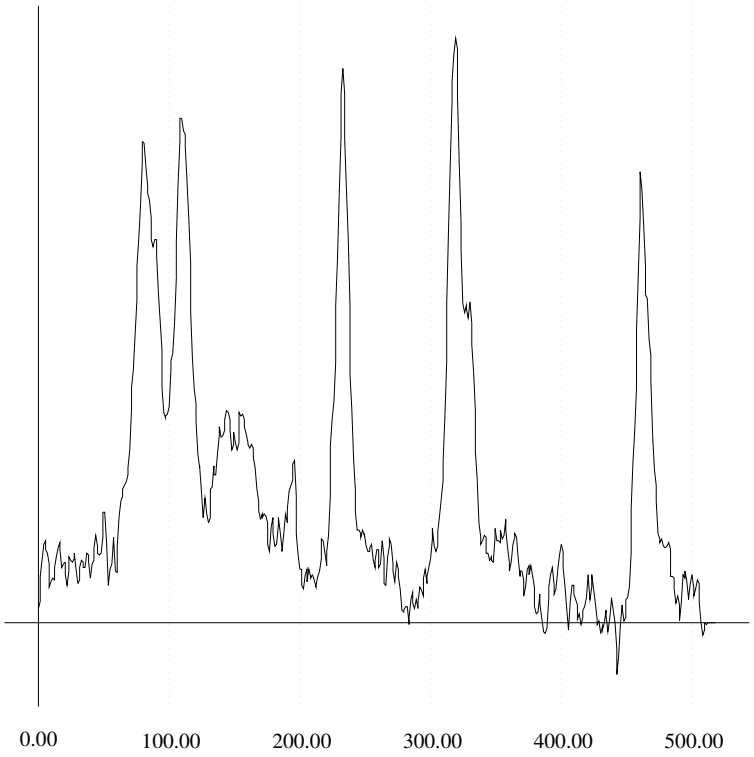

Figure 5.6. Mean spectrum for the spectra of Walker's carcinoma.

For each pair, some spectrum values had a highly significant correlation with spectrum class ($p<0.01$). These were typically in groups of five or six adjacent values, and coincided with the positions of the peaks in the spectra, as would be expected. To see whether a lower line broadening factor might be more effective for this analysis the spectral processing steps were repeated three times for each FID using a line-broadening factor of 5, 10 and 20 Hz. For each different factor 512 datapoints were extracted as before and the correlation matrix were examined. It was found that a factor of 10 Hz produced the the highest set of correlations of the datapoint variables with class. On the basis of these results the datapoints which had been obtained using the factor of 10 Hz were retained and used for all further analysis. The wavelet coefficients were also recalculated using these spectra.

The value with the highest correlation (with tissue type) from each of the groups of highly correlated values was selected and entered into the discriminant program.

Figure 5.7 shows the location of these values on a typical spectrum (whose noisy appearance is due to the low line-broadening factor, chosen for optimal discrimination) and Table 5.5 shows the location and correlation value of the most highly correlated datapoint in each region for each pair of tissue types. As can be seen from Figure 5.7, most of the significantly correlated values correspond to peaks in the spectrum labelled as follows: A, B, C (PME), D, E (unassigned), F(P_i), G(PCr), H(γATP), I,J(αNTP), K(NAD) and L(βNTP). Some of the values, however occurred in the region of the shoulder of a peak [Tate *et al.*, 1996a]. There is no entry in Table 5.5 for liver/brain correlations since these can be easily distinguished

Class	Correlation of class with peaks											
	A	B	C	D	E	F	G	H	I	J	K	L
h9618a & GH3				.82			.7	.82	.74			.6
h9618a & Walker	.65			.88				.6	.68		.56	.71
h9618a & h7777				.77			.75					.7
GH3 & Walker											.6	
liver & hepatomas	.74	.71		.7	.78	.67				.81	.65	
liver & all tumours		.64			.58	.64				.64	.56	
brain & all tumours		.71	.71		.68		.86					

Table 5.5. Table showing absolute correlation coefficients between class and value of datapoint at each of the peaks A to L shown on Figure 5.7. Note, only highly significant correlations ($p<0.01$) are shown.

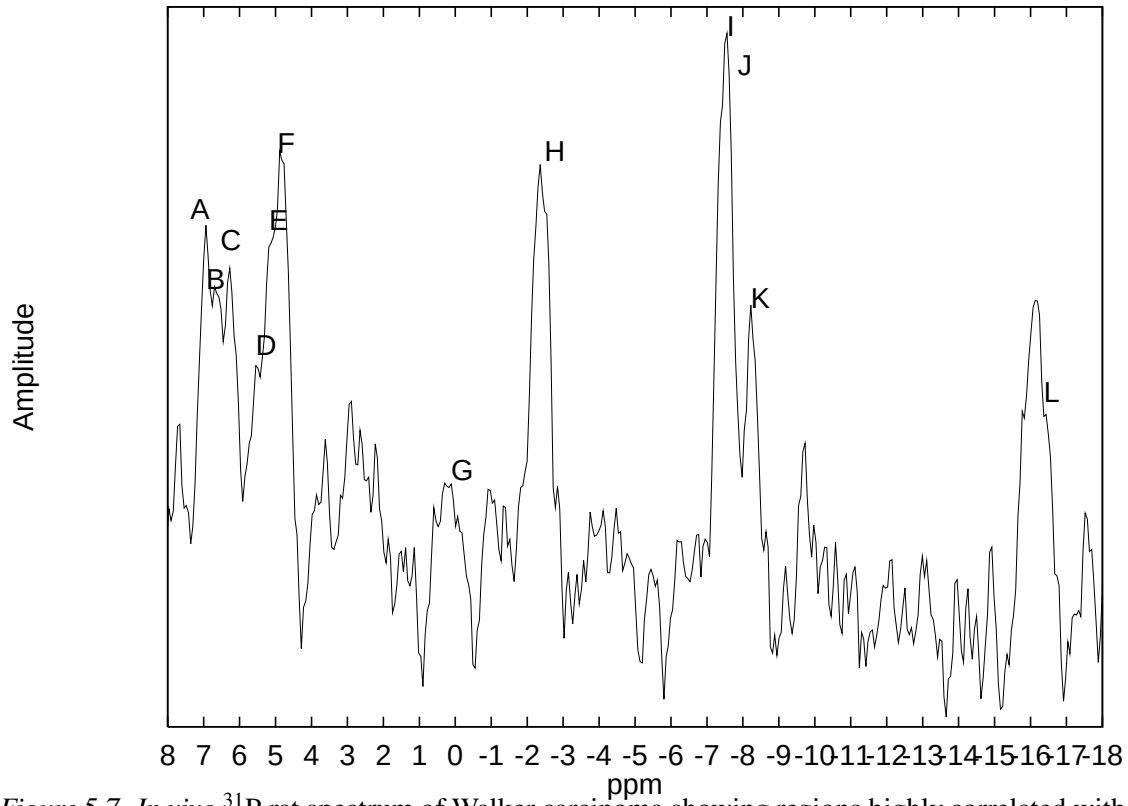

Figure 5.7. In vivo ^{31}P rat spectrum of Walker carcinoma showing regions highly correlated with class, assigned as follows: A, B, C (PME), D, E (unassigned), F(P_i), G(PCr), H(γATP), I,J(αNTP), K(NAD) and L(βNTP).

5.3.4 Classification Results

Table 5.6 summarises the results from discriminant analysis when the leave-one-out method was used to create the test set. Despite the small number of values that could be used in the program, at least 86% of the spectra were assigned correctly for each pair of tissue types. The method used for selecting variables was a form of peak selection. However, unlike the usual method of peak identification and selection, this method used the differences between groups of spectra to identify the important datapoints. I have included the results of the pair h9618a & h7777 hepatomas for completeness although I are aware that their discrimination could be by chance, due to the small number of h7777 hepatomas.

Class	No. correct	Peaks used
h9618a & GH3	20/20 (100%)	D H I
h9618a & Walker	22/23 (96%)	D I L
h9618a & h7777	13/14 (93%)	D G L
GH3 & Walker	18/21 (86%)	H K
liver & tumours	41/45 (91%)	B E J
liver & hepatomas	24/24 (100%)	E A
brain & tumours	45/45 (100%)	B C G

Table 5.6. Classification results when highly correlated datapoint values (from the labelled peak regions) were used in the discriminant program

Class	No. correct
h9618a & GH3	19/20 (95%)
h9618a & Walker	21/23 (91%)
h9618a & h7777	13/14 (93%)
GH3 & Walker	15/21 (71%)
liver & tumours	43/45 (96%)
liver & hepatomas	20/24 (83%)
brain & tumours	40/45 (89%)

Table 5.7. Classification results when wavelet coefficients from the PME region of the spectra were used in the discriminant program

Wavelet Coefficients

For this analysis, 32 points from the PME region (peaks A–C, ppm 7.6–5.73) were selected and wavelet transformed. Since there were only $32 = 2^5$ coefficients representing 5 scale levels only half the coefficients, i.e. the last 16 coefficients numbered from 17-32 were discarded for this analysis. Correlation coefficients were calculated for each of the remaining 16 wavelet coefficients with spectrum class. For most pairs of tissue types a few wavelet coefficients were significantly correlated with class. These were selected for use in the discrimination program. Table 5.7 shows the results. Apart from GH3 & Walker, very good results were obtained (at least 83% correct) for each pair of tissue types, showing that they can be discriminated using only the PME region on the basis of peak shape. Examination of this region by eye showed at least three clear peaks in many of the spectra, but their positions shifted quite considerably from spectrum to spectrum (by as much as 0.2 ppm), perhaps because of differences in pH or in Mg2+ content. An advantage of wavelet transformation is that the wavelet coefficients may not be so dependent on these positions as the original datapoints. This is particularly the case with the first few wavelet coefficients which represent large scale features in the transformed region of the spectrum. I found that I needed to use only the first two wavelet coefficients to discriminate successfully between h9618a & Walker and h9618a & GH3, indicating discrimination was unaffected by exact location of the peaks.

5.3.5 Discussion

In this study the data set was split into pairs of classes and each pair was analysed separately. It was possible to discriminate between the pairs of tissue types with a success rate of at least 86% and regions of importance, consisting of a few data points, could be clearly identified - a fact I found surprising. This meant that it was possible to pinpoint the exact regions in the spectra which were important for discrimination.

The wavelet transform has the advantage that it encodes information concerned with the shape of spectra. Highly correlated wavelet coefficients from the PME region from some pairs of tissue types produced better results than the original datapoints from the same region. This suggests that the ability of the wavelet transformation to discriminate on the basis of peak shape has allowed it to distinguish groups of spectra in which the components of the combined PME peak differ, even though the spectroscopic method used was incapable of resolving them.

These results suggest that lipid precursor signals contain important information for automated cancer diagnosis, with only minimal pre-processing. It also indicated which other region in the spectra might be important for such diagnosis. If the discrimination achieved in this study could be repeated on patients, it could have a clinical application for non-invasive diagnosis or grading of tumours. At present, diagnosis and grading are decided upon mainly by subjective (and labour-intensive) histological examinations of biopsies. If the same information could be obtained from MRS it would obviate the necessity of a biopsy, an important consideration if the tumour is in an inaccessible location, e.g. the brain. The fact that it was possible to discriminate between one tumour type and another was very encouraging, not only because this would be useful in its own right, but also because it indicates that it may be feasible to extend these methods to discriminate between more than two groups. For this to be possible, however, we would need much larger data sets.

5.4 Summary

This chapter describes how the methods described in Chapter 4 were used to classify two sets of *in vivo* data; ^{13}C spectra of subcutaneous fat from a group of 75 subjects of three classes, vegan, vegetarian and omnivore and ^{31}P spectra of 3 classes of normal and 4 classes of tumourous tissue in rats. With the ^{13}C data it was possible to assign 93% of the subjects to their correct class using either peak heights or principal components of the wavelet coefficients as the extracted measurements when the vegetarians were excluded from the analysis and 75% (for peak heights) or 80% (wavelet coefficients) correctly when they were included. For the analysis of the ^{31}P data the classes were considered in pairs, due to the small number of subjects in each group. Here classification rates of between 83% and 100% were achieved depending on which pair of classes were being discriminated. I consider these results very encouraging. While ideally I would have liked success rates of 100% this was not likely to be achievable with either sets of data. For the vegan study the diets of the groups vary widely and the subjects were categorised by their stated diets, which may not have been quite the same as their actual diets. For example one of the vegans (a subject who was consistently misclassified) had admitted to eating chocolate! For the ^{31}P study the number of variables that could be used in the discriminant functions were restricted to only two or three, depending on which class pairs were being analysed. Thus it was possible to use only part of the potentially useful information.

For both data sets these results showed that it was possible to successfully classify most of the spectra, and they demonstrated that fully automated feature extraction was feasible, even when the small size of the data sets meant that very few features could be used in the discriminant program. In addition it was found that, since these data sets could be classified using peak intensities or datapoints which required no further processing, the features could be related back to the original data very easily.

The results of this study also showed that both the DWT and PCA provided very effective methods for reducing the dimensionality of the feature space. Used together they provide a powerful combination.

Chapter 6

Discussion

This thesis describes the investigation of pattern recognition techniques for the analysis of MRS data and develops a fully automated prototype system for classifying *in vivo* spectra. The motivation for the project and relevant theory and background are given in Chapters 1–3. Chapters 4 and 5 then discuss the development of the system and the results of applying this system to two sets of data. The purpose of this final chapter is to sum up the achievements of the work described in this thesis, to discuss its limitations and to offer suggestions as to how the results from the system might be presented.

This chapter is divided as follows:

- Original content
- Discussion of the limitations of the system
- Presentation of Results
- Final Words

6.1 Original Content

The main aim of this research was to investigate and develop automated methods for the analysis and interpretation of *in vivo* MRS data for clinical applications. This involved developing a prototype system for discriminating between different classes of spectra.

The prototype system was developed using two sets of *in vivo* data. The methods that were used to implement this system were described in detail in Chapter 4. Chapter 5 then presented the results of applying these methods to the two data sets, and showed that for most groups of spectra very good classification results were obtained. Using linear discriminant analysis it was possible to classify at least 93% of the individuals from the two main groups of ^{13}C spectra correctly using either peak heights or principal components of the wavelet transformed data. It was also possible to classify correctly many of the data from obtained from the ^{31}P study using either datapoints from the spectrum or wavelet coefficients from a selected region of the spectrum. Good results were obtained for the ^{31}P study despite the fact that the average number of spectra in each class

was only ten, meaning that a maximum of three variables could be used in the linear discriminant program.

These results are very encouraging for two main reasons. Firstly, they show that it is possible to produce a system to classify these data in which all the stages including filtering, phasing, peak alignment, feature extraction and classification is fully automated. To my knowledge this is the first time a classification system has been developed to classify *in vivo* data in which no manual intervention is required for pre–processing the spectra. Secondly, it showed that the spectra could be classified using information from the whole spectrum, without the need for identifying or explicitly quantifying the peaks. This confirms the results of a number of other studies of *in vitro* data, for example [Howells *et al.*, 1992b] [Somorjai *et al.*, 1995b].

This thesis extends previous studies by carrying out a thorough investigation of methods for extracting features from *in vivo* spectra which use no prior knowledge of biochemistry. A number of methods are investigated for extracting salient features from the spectra including principal component analysis, wavelet analysis and correlation methods. The fact that these methods could be successfully used to classify the two sets of data studied here indicates that they may be more widely applicable to other data sets.

A somewhat surprising result was that it was possible to classify the spectra from both data sets using features that required very little processing of the data, i.e. peak heights and spectral datapoints. These advantage of using such features is that they can be easily related back to the original spectrum and can thus pinpoint the exact regions of the spectra that are important for discrimination.

In this study it was possible to discriminate between the spectra using features that can be related directly back to the peaks in the spectra, and thus to specific metabolites. Whether or not this will be possible with spectra acquired under less ideal conditions still remains to be seen. In this case it may be necessary to use more powerful techniques such as the DWT and PCA for preprocessing and feature extraction. This study showed that the DWT proves to be a useful tool for classifying both sets of spectra. Unfortunately was not possible to investigate properly the use of PCA for the ^{31}P data sets as the numbers in each class were too small. However, this method could be used to reduce considerably the number of features in the ^{13}C data set. Particularly good classification results were obtained when the PCA and the DWT were combined. This is because, while both methods allow the data to be expressed more succinctly they do so in different ways. The first few PC's represent the variation between the samples whereas the first few wavelet coefficients will represent the variation within the sample. The two methods can thus be combined very advantageously.

The results using the wavelet transform showed that the spectra could be classified completely automatically with no need for the identification or quantification of peaks. These results suggest that the wavelet transform might prove a useful tool for classifying groups of spectra that can be discriminated on the basis of line-shape or spectral pattern.

In order to assess the real worth of these techniques as a general tool for clinical MRS, it will be necessary to see how successful they are when applied to other sets of data.

6.2 Limitations of the System

This study, as is common with many studies of medical data, was limited by lack of data. Ideally I would have liked to have had much larger datasets, with the order of hundreds rather than tens of subjects in each category. I would also have liked to have other data sets on which to test the

methods that were developed.

The small numbers of data restricted this study in a number of ways. Firstly, it ruled out the possibility of investigating other classification methods such as non–parametric discriminant analysis. While LDA produced reasonable classification results for the two sets of data in this study, other methods may prove to be more suitable for a real system.

Secondly the small number of samples severely restricted the number of variables that could be used in the discriminant program. Most authors advise that, in order to avoid the problem of overfitting, the number of variables that are used in the discriminant program should be should be no more than one third the number of samples in the smallest class [Massart *et al.*, 1988] [Kowalski and Wold, 1982]. This meant that for the ^{31}P study a maximum of three variables could be used in the program and therefore all the potential information could not be fully utilised. The small numbers of data also meant that it was not possible to fully investigate the use of PCA for feature reduction. This is because PCA can be unstable if there is a large number of variables compared with the number of samples.

A potential problem of using correlation methods for selecting features is that it introduces the possibility of selection bias. When the sample size is small it is necessary to use the test set both for selecting the variables and for testing the classification rule. Selection bias was not a major worry for the ^{13}C data, because the number of samples was relatively large and it was possible to select features using only half the data (set 1). Selection bias could also be prevented by choosing the wavelet coefficients on the basis of scale levels and the PC's on the basis of the variance that they explained.

Selection bias was a worry for the ^{31}P data set because with this data it was necessary to use the test set to select the features. Although features were only selected if it appeared that there was a good biochemical reason for doing so, i.e. if they corresponded to peak regions in the spectra, which were known to be related to biochemical differences between spectra, any conclusions drawn from this study would need to be confirmed using more data.

Another limitation of the work described in this thesis is that it does not address the problem of how the classification results should be presented. For a system that will be used for real clinical applications this will be an important consideration. The following section provides a few ideas of how the results of the system might be presented.

6.3 Presentation of Results

Although the development of a classification system for a particular set of data may be a very useful exercise in its own right, the main purpose of this study was not specifically to investigate biochemical differences between different classes of spectra, but to provide methods which will help with the interpretation of MRS data in order that they can be used as a clinical tool. The decision to investigate pattern recognition methods of analysis and to develop a prototype classification system was taken because most clinical applications will require reliable categorisation of spectra into distinct groups according to the disease state or stage of the tissue being examined.

Thus it is important to consider how a classification system might actually be used in practice to help the decision making process of the clinician, and to consider what the output from the system should be. This means considering what sort of information should be provided, and how the results of the analysis should be presented in a format that is accessible to the clinician.

The most basic requirement for the system are that it will take an FID signal (or spectrum)

as input and produces as output a classification for the signal, together with the probability of the individual belonging to a particular class.

However, because the system is intended to be used as a clinical aid, it is also important that it should also provide as much information as possible as to how any decisions were reached. Thus, the system should provide not only the results of the classification, but should also provide reasons why a subject has been assigned to a particular class.

Since the decision to assign a certain individual to a particular class will be determined by its linear discriminant score or scores, it would be useful to give the value of this score and show how this can be compared with the scores of individuals whose class is known. A plot of these scores will show how the discriminant score of the subject compares with others of the same class, and can show how typical or atypical the score is for that class. If there are only two classes, and thus only one discriminant function, this can be achieved by using a one–dimensional dot plot, or alternatively a histogram. If there are more than two classes, a scatterplot can be used.

The linear discriminant functions, and therefore the discriminant scores, will very much depend on the variables that have been selected for use in the discriminant program. If these features can be related to meaningful factors in the spectrum, such as relative levels of certain metabolites, it may be useful to output the values of these features. A method of presenting results commonly used in clinical chemistry is to present such results in a table. One row of the table includes the mean values of all metabolites measured and the other row contains the values for the sample. Another method would be to display the values of the features, together with those of a representative sample of the training data two at a time on a scatterplot.

What constitutes meaningful features will depend very much on the particular application and will most probably need to be determined by those with an expert knowledge of the biochemistry of the spectra. The closer these features are to the actual peaks in the spectra, the easier it will be to give them biochemical meaning. When the features are peak intensities, or datapoints, biochemical interpretation will normally be possible by assigning the peak or ppm region to a known metabolite. When the features are wavelet coefficients however, such assignment will be more difficult. Although the index of the wavelet coefficients will give an idea of the scales and position of the features in the spectrum that they represent, it may be more difficult to relate the wavelet coefficients back to individual metabolites. It may be even more difficult to interpret principal components. Although in some cases it may be possible to ascribe some physical meaning to individual principal components, this can only be hoped for rather than expected.

Suggested output for the classification system at each of the three stages in the system are as follows.

Stage 1 Spectral Processing

- A plot of the spectrum to check for major errors in data acquisition or processing (for example incorrect peak alignment).
- A plot of the spectrum on a overlay plot with mean spectra for the classes to see how the spectrum compares. It may be possible at this stage to see which group it belongs to.

Stage 2 Feature Extraction

- Scatterplots of the features to be used in the discriminant program.
- A table of values of these features

Stage 2 Classification

- The classification for the spectrum, together with the probabilities that the spectrum belong to a certain class
- A 1–dimensional dot plot or 2–dimensional scatterplot of the discriminant scores of the spectrum together with those of the training data.

6.4 Conclusions

This research has developed a novel and useful technique for analysing MRS data. The main objective of this thesis was to develop automated methods for analysing and interpreting *in vivo* MRS data. The approach used was somewhat different to most previous studies for automated analysis, in that the aim was to provide qualitative, rather than quantitative, information, that is to identify the type of tissue from which the spectrum was derived. The standard approach is to identify the metabolites and then to look for differences. In this thesis I have shown that it is possible to directly classify tissue samples without first evaluating the metabolite concentrations. Metabolite measurements require prior knowledge of their significance and can also be difficult to obtain. This thesis has demonstrated that a statistical approach to identifying significant features is sufficient.

A major by-product of this approach is we can derive quantitative information about the metabolite concentrations. Additionally, the identification of these discriminatory features may provide important clues for subsequent biochemical analysis [Tate *et al.*, 1996a].

This approach is novel and useful because the information is suitable for direct use by clinicians. It is hoped that this research will form a suitable basis for direct clinical application of MRS and will help facilitate its use as a clinical tool.

Bibliography

[Abbott, 1994] Vikrant Abbott. Nuclear magnetic resonance – peak detection and area estimation. Final year project report, University of Sussex, School of Cognitive and Computing Sciences, April 1994.

[Aeberhard *et al.*, 1994] Stefan Aeberhard, Danny Coomans, and Olivier De Vel. Comparative analysis of statistical pattern recognition methods in a high dimensional setting. *Pattern Recognition*, 27(8):1065–1077, 1994.

[Andrew *et al.*, 1990] E. Raymond Andrew, Graeme Bydder, John Griffiths, and Peter Styles. *Clinical Magnetic Resonance Imaging and Spectroscopy*. John Wiley and Sons, 1990.

[Anthony *et al.*, 1994] M. L. Anthony, B. C. Sweatman, C. R. Beddell, J. C. Lindon, and J. K. Nicholson. Pattern recognition classification of the site of nephrotoxicity based on metabolic data derived from proton nuclear magnetic resonance spectra of urine. *Molecular Pharmacology*, 46(1):199–211, 1994.

[Beckmann, 1992] N. Beckmann. *In Vivo* ^{13}C spectroscopy in humans. In *NMR Basic Principles and Progress, Volume 28*. Springer-Verlag, Berlin, Heidelberg, 1992.

[Bellman, 1957] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[Bishop, 1995] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[Bock, 1994] J. L. Bock. NMR in clinical chemistry – Where do we stand? *Clinical Chemistry*, 40(7):1215–1217, 1994.

[Bos and Vrielink, 1994] M. Bos and J. A. M. Vrielink. The wavelet transform for preprocessing IR-spectra in the identification of monosubstituted and disubstituted benzenes. *Chemometrics and Intelligent Laboratory Systems*, 23(1):115–122, 1994.

[Bottomley, 1989] P. Bottomley. Human *in vivo* NMR spectroscopy in diagnostic medicine: Clinical tool or research probe? *Radiology*, 170(1):1–15, January 1989.

[Bracewell, 1986] Ronald N. Bracewell. *The Fourier Transform and its Applications*. McGraw-Hill, New York, 1986.

[Breiman *et al.*, 1984] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[Cady, 1990] Ernest Cady. *Clinical Magnetic Resonance Spectroscopy*. Plenum Press, 1990.

[Chatfield and Collins, 1980] C. Chatfield and A. J. Collins. *Introduction to Multivariate Analysis*. Chapman & Hall, 1980.

[Chen and Kan, 1988] Ching-Nien Chen and Lou-Sing Kan. An iterative phase correction program for nuclear magnetic resonance (NMR) spectra. *Computer Methods and Programs in Biomedicine*, 26(1):81–84, 1988.

[Confort-Gouny *et al.*, 1993] S. Confort-Gouny, J. Viondury, F. Nicoli, P. Dano, A. Donnet, N. Grazziani, J. L. Gastaut, F. Grisoli, and P. J. Cozzone. A multiparametric data analysis showing the potential of localized proton MR spectroscopy of the brain in the metabolic characterization of neurological diseases. *Journal of the Neurological Sciences*, 118(2):123–133, 1993.

[Coomans and Broeckaert, 1986] D. Coomans and I. Broeckaert. *Potential Pattern Recognition in Chemical and Decision Making*. Research Studies Press Ltd., John Wiley and Sons Inc., 1986.

[Daubechies, 1992] Ingrid Daubechies. *Ten Lectures on Wavelets*. CBMS Regional Conference Series in Applied Mathematics, 1992.

[de Beer and van Ormondt, 1992] R. de Beer and D. van Ormondt. Fitting procedures. In *NMR Basic Principles and Progress*, volume 26. Springer Verlag, 1992.

[de Certaines *et al.*, 1992] J. D. de Certaines, W. M. M. J. Bovée, and F. Podo, editors. *Magnetic Resonance Spectroscopy in Biology and Medicine*. Pergamon Press, 1992.

[Diop *et al.*, 1992] A. Diop, A. Briguet, and D. Graveron-Demilly. Automatic *in vivo* NMR data processing based on an enhancement procedure and linear prediction method. *Magnetic Resonance in Medicine*, 27:318–328, 1992.

[Dolenko and Somorjai, 1995] B. Dolenko and R. L. Somorjai. Time well spent: Preprocessing of MR spectra for greater classification accuracy. In *Society for Magnetic Resonance in Medicine 1995 meeting*, page 1936, August, 1995.

[Duda and Hart, 1973] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.

[El-Deredy and Branston, 1994] W. El-Deredy and N. M. Branston. Comparison of k-NN and backpropagation in classifying 1H NMR tumour spectra. In *Proceedings of the 2nd International Conference on Artificial Intelligence Applications*, pages 473–480, Cairo, 1994.

[El-Deredy *et al.*, 1995] W. El-Deredy, N. M. Branston, A. A. Sankar, J. L. Darling, S. R. Williams, and D. G. T. Thomas. Identification of metabolites in proton NMR tumour spectra using artificial neural networks. *British Journal of Neurosurgery*, 9:255, 1995.

[Everitt and Dunn, 1991] B. S. Everitt and G. Dunn. *Applied Multivariate Data Analysis*. Edward Arnold, 1991.

[Fisher, 1936] R. A. Fisher. The use of multiple measurements on taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[Friedman and Tukey, 1974] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions in Computing*, 23:881–889, 1974.

[Friesen *et al.*, 1995] L. Friesen, G. Scarth, B. Dolenko, A. Nikulin, N. Pizzi, R. L. Somorjai, F. Guijon, M. Paraskevas, and I. Smith. Dysplasia of the uterine cervix: Classification via ^{1}H MRS and multivariate analysis. In *Society for Magnetic Resonance in Medicine 1995 meeting*, August 1995.

[Fukunaga, 1990] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, Boston, second edition, 1990.

[Gadian, 1982] David G. Gadian. *Nuclear Magnetic Resonance and its Applications to Living Systems*. Oxford: Clarendon Press, 1982.

[Gadian, 1995] David G. Gadian. *Nuclear Magnetic Resonance and its Applications to Living Systems*. Oxford: Clarendon Press, second edition, 1995.

[Gartland *et al.*, 1991] K. P. R. Gartland, C. R. Beddell, J. C. Lindon, and J. K. Nicholson. Application of pattern recognition methods to the analysis and classification of toxicological data derived from proton nuclear resonance spectroscopy of urine. *Molecular Pharmacology*, 39(5):629–642, 1991.

[Glover and Hopke, 1992] D. M. Glover and P. K. Hopke. Exploration of multivariate chemical data by projection pursuit. *Chemometrics and Intelligent Laboratory Systems*, 16(1):45–59, 1992.

[Hagberg *et al.*, 1995] Gisela Hagberg, Allessandro P. Burlina, Irina Mader, Werner Roser, Ernst W. Radue, and Joachim Seelig. *In vivo* proton MR spectroscopy of human gliomas: Definition of metabolic coordinates for multi-dimensional classification. *Magnetic Resonance in Medicine*, 34(2):242–252, August 1995.

[Heerschap *et al.*, 1996] A. Heerschap, G. Jager, J. Barentsz, A. de Koster, J. de la Rosette, G. Oosterhof, F. Debruyne, and J. Ruijs. ^{1}H magnetic resonance of localized prostate cancer. Submitted for publication, 1996.

[Hennel and Klinowski, 1993] J. W. Hennel and J. Klinowski. *Fundamentals of Nuclear Magnetic Resonance*. Longman Scientific & Technical, Hong Kong, 1993.

[Holmes *et al.*, 1994] E. Holmes, P. J. D. Foxall, J. K. Nicholson, G. H. Neild, S. M. Brown, C. R. Beddell, B. C. Sweatman, E. Rahr, J. C. Lindon, M. Spraul, and P. Neidig. Automatic data reduction and pattern recognition methods for analysis of ^{1}H nuclear magnetic resonance spectra of human urine from normal and pathological states. *Analytical Biochemistry*, 220(2):284–296, 1994.

[Howells *et al.*, 1992a] S. L. Howells, R. J. Maxwell, and J. R. Griffiths. Classification of tumour ^{1}H NMR spectra by pattern recognition. *NMR in Biomedicine*, 5:59–64, 1992.

[Howells *et al.*, 1992b] S. L. Howells, R. J. Maxwell, A. C. Peet, and J. R. Griffiths. An investigation of tumor ^{1}H nuclear magnetic resonance spectra by the application of chemometric techniques. *Magnetic Resonance in Medicine*, 28:214–236, 1992.

[Howells *et al.*, 1993a] S. L. Howells, R. J. Maxwell, F. A. Howe, A. C. Peet, M. Stubbs, L. M. Rodrigues, S. P. Robinson, S. Baluch, and J. R. Griffiths. Pattern recognition of ^{31}P magnetic resonance spectroscopy tumour spectra obtained *in vivo*. *NMR in Biomedicine*, 6:237–241, 1993.

[Howells *et al.*, 1993b] S. L. Howells, R. J. Maxwell, L. M. Rodrigues, S. J. Crabb, A. C. Peet, and J. R. Griffiths. Pattern recognition of ^{31}P NMR spectra of tumours *in vivo*. In *Proceedings of the 12th annual meeting of the SMRM*, page 1002. Society for Magnetic Resonance in Medicine, August 1993.

[James, 1985] M. James. *Classification Algorithms*. Collins, London, 1985.

[Joliot *et al.*, 1991] M. Joliot, B. M. Mazoyer, and R. H. Huesman. *In Vivo* NMR spectral parameter estimation: A comparison between time and frequency domain methods. *Magnetic Resonance in Medicine*, 18:358–370, 1991.

[Kormos and Waugh, 1983] D. W. Kormos and J. S. Waugh. Abstract factor analysis of solid-state nuclear magnetic resonance spectra. *Analytical Chemistry*, 55(4):633–638, 1983.

[Kowalski and Bender, 1976] B. R. Kowalski and C. F. Bender. An orthogonal feature selection method. *Pattern Recognition*, 8:1–4, 1976.

[Kowalski and Reilly, 1971] B. R. Kowalski and C. A. Reilly. Nuclear magnetic resonance spectral interpretation by pattern recognition. *Journal of Physical Chemistry*, 75:1402–1411, 1971.

[Kowalski and Wold, 1982] Bruce R. Kowalski and Svante Wold. Pattern recognition in chemistry. In P. R. Krishnaiah and L. N. Kanal, editors, *Handbook of Statistics*, volume 2, pages 673–697. North Holland Publishing Company, 1982.

[Krzanowski, 1988] W. J. Krzanowski. *Principles of Multivariate Analysis*. Clarendon Press, Oxford, 1988.

[Leach, 1992] M. O. Leach. Spatially localized nuclear magnetic resonance. In S. Webb, editor, *The Physics of Medical Imaging*, chapter 8, pages 389–487. Institute of Physics Publishing, Bristol, 3rd edition, 1992.

[Malinowski and Howery, 1980] E. R. Malinowski and D. G. Howery. *Factor Analysis in Chemistry*. John Wiley and Sons, 1980.

[Mallat and Zhong, 1992] Stephane Mallat and Sifen Zhong. Characterization of signals from multiscale edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(7), July 1992.

[Mallat, 1989] S. G. Mallat. A theory for multiresolution signal decomposition – the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

[Massart *et al.*, 1988] D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte, and L. Kaufman. *Chemometrics: A Textbook.* Elsevier Science Publishers, 1988.

[Masters, 1995] Timothy Masters. *Advanced Algorithms for Neural Networks – A C++ Sourcebook.* John Wiley and Sons, 1995.

[McLachlan, 1992] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York, 1992.

[Miclet, 1986] Lauren Miclet. *Structural Methods in Pattern Recognition.* North Oxford Academic Publishers Ltd., 1986.

[Miller, 1990] A. J. Miller. *Subset Selection in Regression.* Chapman & Hall, 1990.

[Negendank, 1992] William Negendank. Studies of human tumours by MRS: A review. *NMR in Biomedicine*, 5:303–324, 1992.

[Nikulin *et al.*, 1995] A. Nikulin, K. M. Briere, L. Friesen, I. C. P. Smith, and R.L. Somorjai. Genetic algorithm-guided optimal attribute selection: A novel preprocessor for classifying MR spectra. In *Society for Magnetic Resonance in Medicine 1995 meeting*, page 1940, August 1995.

[Norusis, 1994] M. J. Norusis. *S.P.S.S. Advanced Statistics : Release 6.1.* SPSS Inc., Chicago, 1994.

[Peden *et al.*, 1993] C. J. Peden, M. A. Rutherford, J. Sargentoni, I. J. Cox, D. J. Bryant, and L. M. S. Dubowitz. Proton spectroscopy of the neonatal brain following hypoxic-ischemic injury. *Developmental Medicine and Child Neurology*, 35(6):502–510, 1993.

[Press *et al.*, 1992] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, second edition, 1992.

[Preul *et al.*, 1994] M. C. Preul, Z. Caramanos, D. L. Collins, J-G. Villemure, W. Feindel, and D. L. Arnold. Linear discriminant analysis based on proton MR spectroscopic imaging of human brain tumours improves pre-operative diagnosis. In *Proceedings of the 2nd Meeting of the SMR*, page 125. Society of Magnetic Resonance, August 1994.

[Provencher, 1993] Stephen W. Provencher. Estimation of metabolite concentrations from localized *in vivo* proton NMR spectra. *Magnetic Resonance in Medicine*, 30:672–679, 1993.

[Ross and Michaelis, 1994] B. Ross and T. Michaelis. Clinical applications of magnetic resonance spectroscopy. *Magnetic Resonance Quarterly*, 10(4):191–247, 1994.

[Saeed and Menon, 1993] N. Saeed and D. K. Menon. A knowledge-based approach to minimize base line roll in chemical shift imaging. *Magnetic Resonance in Medicine*, 29(5):591–598, 1993.

[Saito, 1994] Naoki Saito. *Local Feature Extraction and its Applications Using a Library of Bases*. PhD thesis, Yale University, 1994.

[Sanders and Hunter, 1993] Jeremy K. M. Sanders and Brian K. Hunter. *Modern NMR Spectroscopy: A Guide for Chemists*. Oxford University Press, 1993.

[Schalkoff, 1992] Robert J. Schalkoff. *Pattern Recognition: Statistical, Structural and Neural Approaches*. John Wiley and Sons, 1992.

[Scott, 1992] D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, New York, 1992.

[Serrai *et al.*, 1995] H. Serrai, L. Senhadji, N. Le Tallec, E. Le Rumeur, and J. D. De Certaines. Quantification of mrs time-domain signal using wavelet transform application to automatic processing of serial ^{31}P MRS acquisition on working rat muscle spectra of tumours in rats using statistical pattern recognition. In *Society for Magnetic Resonance in Medicine 1995 Meeting*, page 1954, August 1995.

[Siegel and Castellan, 1988] Sidney Siegel and John N. Castellan. *Nonparametric Statistics for the Behavioural Sciences*. McGraw Hill, second edition, 1988.

[Silverman, 1986] B. W. Silverman. *Density Estimation*. Chapman & Hall, 1986.

[Sjostrom and Kowalski, 1979] Michael Sjostrom and Bruce R. Kowalski. A comparison of five pattern recognition methods based on the classification results from six real data bases. *Analytica Chimica Acta*, 112:11–30, 1979. Computer Techniques and Optimization, Elsevier Scientific Publishing Company.

[Slichter, 1989] C. P. Slichter. *Principles of Magnetic Resonance*. Springer Series in Solid-State Sciences 1. Springer-Verlag, Berlin, New York, 3rd enl. and updated edition, 1989.

[Somorjai *et al.*, 1995a] R. L. Somorjai, D. J. Kitchen, B. Dolenko, A. Nikulin, G. Scarth, D. Ende, R. Newland, P. Russel, C. E. Mountford, T. Bezabeh, K. M. Briere, C. N. Bernstein, N. M. Pettigrew, K. J. Lewin, and I. C. P. Smith. When all else fails: Multivariate analysis for the robust accurate classification of ^{1}H spectra of colorectal biopsies. In *Society for Magnetic Resonance in Medicine 1995 meeting*, page 1938, August 1995.

[Somorjai *et al.*, 1995b] Ray L. Somorjai, Alexander E. Nikulin, Nic Pizzi, Dick Jackson, Gordon Scarth, Brion Dolenko, Heather Gordon, Peter Russell, Cynthia L. Lean, Leigh Delbridge, Carolyn E. Mountford, and Ian C. P. Smith. Computerized consensus diagnosis: A classification strategy for the robust analysis of MR spectra. 1. Application to ^{1}H spectra of thyroid neoplasms. *Magnetic Resonance in Medicine*, 33:257–263, 1995.

[Spisni, 1992] Alberto Spisni. 1D spectrum analysis. In J. D. de Certaines, W. M. M. J. Bovée, and F. Podo, editors, *Magnetic Resonance Spectroscopy in Biology and Medicine*. Pergamon Press, 1992.

[SPSS Inc., 1987] SPSS Inc. *SPSSX User's Guide*. SPSS Inc., Chicago, 3rd edition, 1987.

[Stoyanova *et al.*, 1995] R. Stoyanova, A. C. Kuesel, and T. R. Brown. Application of principal component analysis for NMR spectral quantitation. *Journal of Magnetic Resonance Series A*, 115(2):265–269, 1995.

[Tate *et al.*, 1995] Rosemary Tate, Des Watson, and Stephen Eglen. Using wavelets for classifying human *in vivo* magnetic resonance spectra. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, volume 103 of *Lecture Notes in Statistics*, pages 377–383. Springer-Verlag, 1995.

[Tate *et al.*, 1996a] A. Rosemary Tate, Simon Crabb, John R. Griffiths, Sian L. Howells, Roy A. Mazucco, Loreta M. Rodrigues, and D. Watson. Lipid metabolite peaks in pattern recognition analysis of tumour *in vivo* MR spectra. *Anticancer Research*, 16(3):1575–1580, 1996.

[Tate *et al.*, 1996b] A. Rosemary Tate, Des Watson, Stephen Eglen, Theodoros N. Arvanitis, E. Louise Thomas, and Jimmy D. Bell. Automated feature extraction for the classification of human *in vivo* ^{13}C NMR spectra using statistical pattern recognition and wavelets. *Magnetic Resonance in Medicine*, 35:834–840, June 1996.

[Thomas *et al.*, 1995] E. L. Thomas, J. D. Hanrahan, J. Sargentoni, D. Azzopardi, M. L. Barnard, D. J. Bryant, J. E. Schwieso, S. R. Bloom, and J. D. Bell. Characterisation of neonatal adipose tissue by *in vivo* ^{13}C magnetic resonance spectroscopy. Submitted for publication, 1995.

[Titterington *et al.*, 1981] D. M. Titterington, G. D. Murray, L. S. Murray, D. J. Spiegelhalter, A. M. Skene, J. D. F. Habbema, and G. J. Gelpke. Comparison of discrimination techniques applied to a complex data set of head injured patients. *Journal of the Royal Statistical Society, Series A – General*, 144(P2):145–175, 1981.

[Tufte, 1983] Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1983.

[van Dijk *et al.*, 1992] J. E. van Dijk, A. F. Mehlkopf, D. van Ormondt, and M. J. Bovee. Determination of concentrations by time domain fitting of proton NMR echo signals using prior knowledge. *Magnetic Resonance in Medicine*, 27:76–96, 1992.

[Vine, 1990] W. Vine. Clinical diagnosis by nuclear magnetic resonance spectroscopy – if not now, when? *Archives of Pathology and Laboratory Medicine*, 114(5):453–462, 1990.

[Wang, 1992] Liqun Wang. *Towards Automatic and Quantitative Processing of Magnetic Resonance Spectroscopy*. PhD thesis, University of Sussex, 1992.

[Weiner, 1988] M. W. Weiner. The promise of magnetic resonance spectroscopy for medical diagnosis. *Invest. Radiol.*, 23:253–261, 1988.

[Weiner, 1994] Michael W. Weiner. Clinical applications of MR spectroscopy and spectroscopic imaging. In *Proceedings of the 2nd Annual Meeting of the SMR*, page 185, August 1994.

[Wu *et al.*, 1995] W. Wu, B. Walczak, D. L. Massart, K. A. Prebble, and I. R. Last. Spectral transformation and wavelength selection in near-infrared spectra classification. *Analytica Chimica Acta*, 315(3):243–255, 1995.

[Wunsch and Laine, 1995] Patrick Wunsch and Andrew Laine. Wavelet descriptors for multiresolution recognition of handprinted characters. *Pattern Recognition*, 28(8):1237–1249, 1995.