Pattern Recognition Analysis of *In Vivo* Magnetic Resonance Spectra

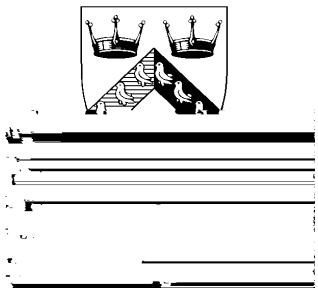
Anne Rosemary Tate

432

Se te ber, 1996

ISSN 1350–3162

UNIVERSITY OF



Cognitive Science

Research Papers

S F

F F

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.

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 ob02(b)-4.11137(e)-279.237(e)-303

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 [

- 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.

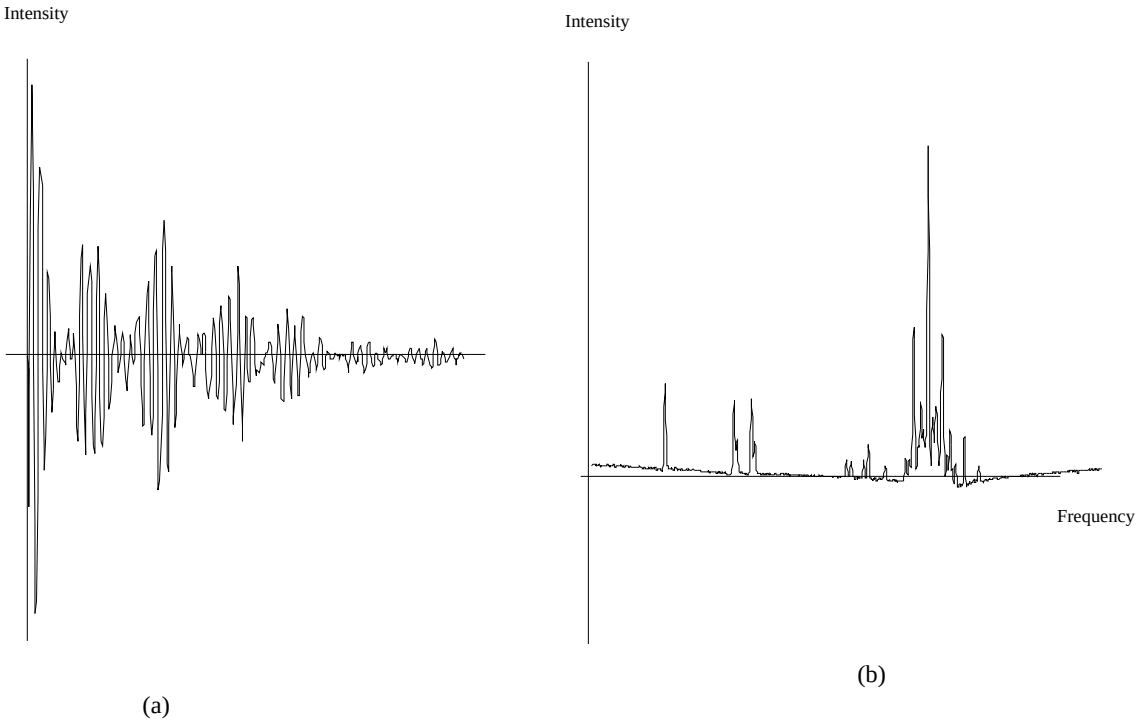


Figure 2.1. ^{13}C FID signal (a) and spectrum (b) from a human thigh

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

narrow signals from metabolites which have a high degree of m

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

$$\begin{aligned} \boldsymbol{r}_2 &= -\boldsymbol{i}_1 \boldsymbol{sin}\theta + \boldsymbol{r}_1 \boldsymbol{cos}\theta \\ \boldsymbol{i}_2 &= \boldsymbol{r}_1 \boldsymbol{sin}\theta + \end{aligned} \tag{2.4}$$

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

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

5

5 *in vivo*

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 $\boldsymbol{B}_0$, 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 fromund

he s t eot in vi

]ersible torsee t ae ofe NMnna

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

S

: *in vivo*

‘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 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 t

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.

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].

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].

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].

'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 maximu'Cistis of d

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.

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,

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

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

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.

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 $\boldsymbol{k}$-nearest neighbour method, the $\boldsymbol{k}$ nearest samples are selected and the object is allocated to the class to which the majority of the $\boldsymbol{k}$ 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 $\boldsymbol{K}$ which satisfies the condition $\int \boldsymbol{K}(\boldsymbol{x})\boldsymbol{dx} = 1$. Usually (but not always) $\boldsymbol{K}$ will be a symmetric probability density function such as the normal density. The kernel estimator with density $\boldsymbol{K}$ is defined by

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

where $\boldsymbol{h}$ is the ***window width***, $\boldsymbol{m}$ is the number of samples and $\boldsymbol{X}_1 \ldots \boldsymbol{X_m}$ are the values for each sample [Silverman, 1986]. $\boldsymbol{K}$ acts as the shape parameter and $\boldsymbol{h}$ is a smoothing parameter. For most applications $\boldsymbol{K}$ is fixed while $\boldsymbol{h}$ 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 $\boldsymbol{h}$ 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

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

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. Th-6.93404(t)-6.93181(n)-4.1097(e)5.64366(c)4.118(T)]Tlle

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

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 subspace

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 $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots . \boldsymbol{x}_{\boldsymbol{n}}]$ can be 'explained' in terms of a smaller number of unobservable latent variables or factors $\boldsymbol{f}_1, \ldots \boldsymbol{f}_{\boldsymbol{k}}$ where $\boldsymbol{k} < \boldsymbol{n}$. The factor model is given by

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

function, an the lare scale (low frequency) fea-
has the disadv ntage that the frequency information
ecause its bas
ures required. This problem may beartially
s whereb the signa to be anal-

3.5. Statistical PR: Appropriate Methods for Analysis of MR

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.

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.

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

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.

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.

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 inclu

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 expone

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

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

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 itF

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 proble

- 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!

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 fewerf sndur dsk(s)-5.52048(o)-4.1856(W)1.53396(cp)-4.10914(p)-4.1091(h)-4.10914(e)-2537(f)-214.9373

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

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.

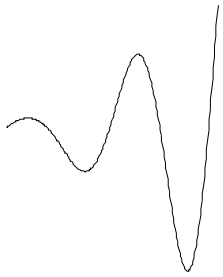
Three types of spectral features were investigated in this research. These features were chosen on the basis of the preliminary investigations:

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 [

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,



$\frac{2^6-2^5}{2}$) 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.

5

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 bef

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

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.

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

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]

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.

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

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

5

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.

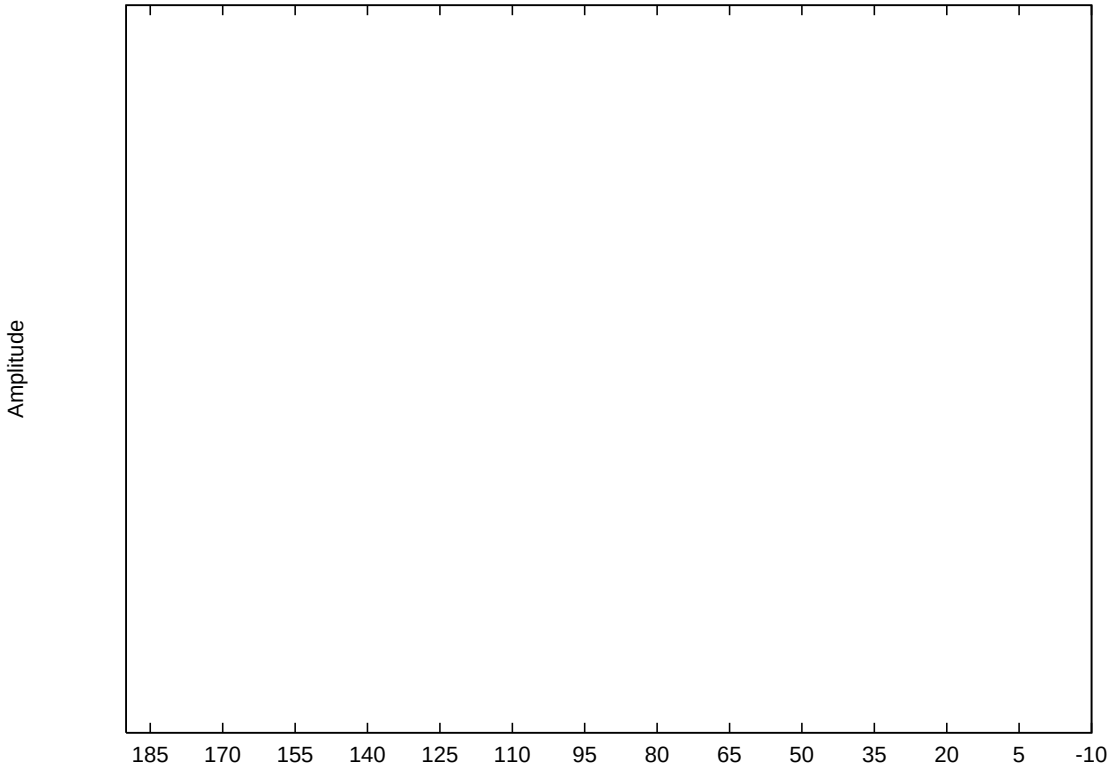
The two specific aims of the work described in this chapter were:

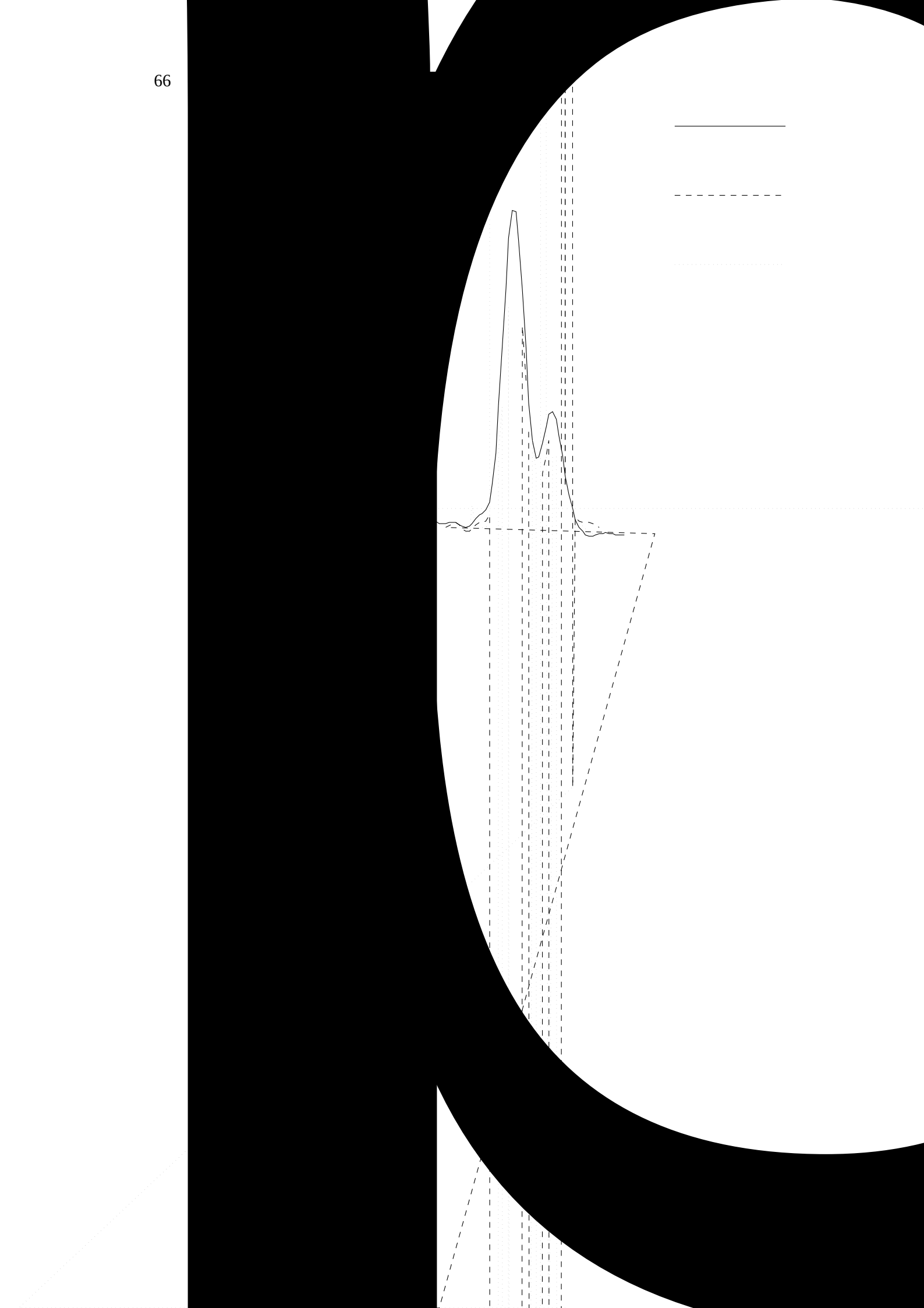
1. to see whether it was possible to design a fully automated system for classifying these spectra and
2. to develop and test methods for automatically extracting features which could be used to discriminate between the different classes of spectra.

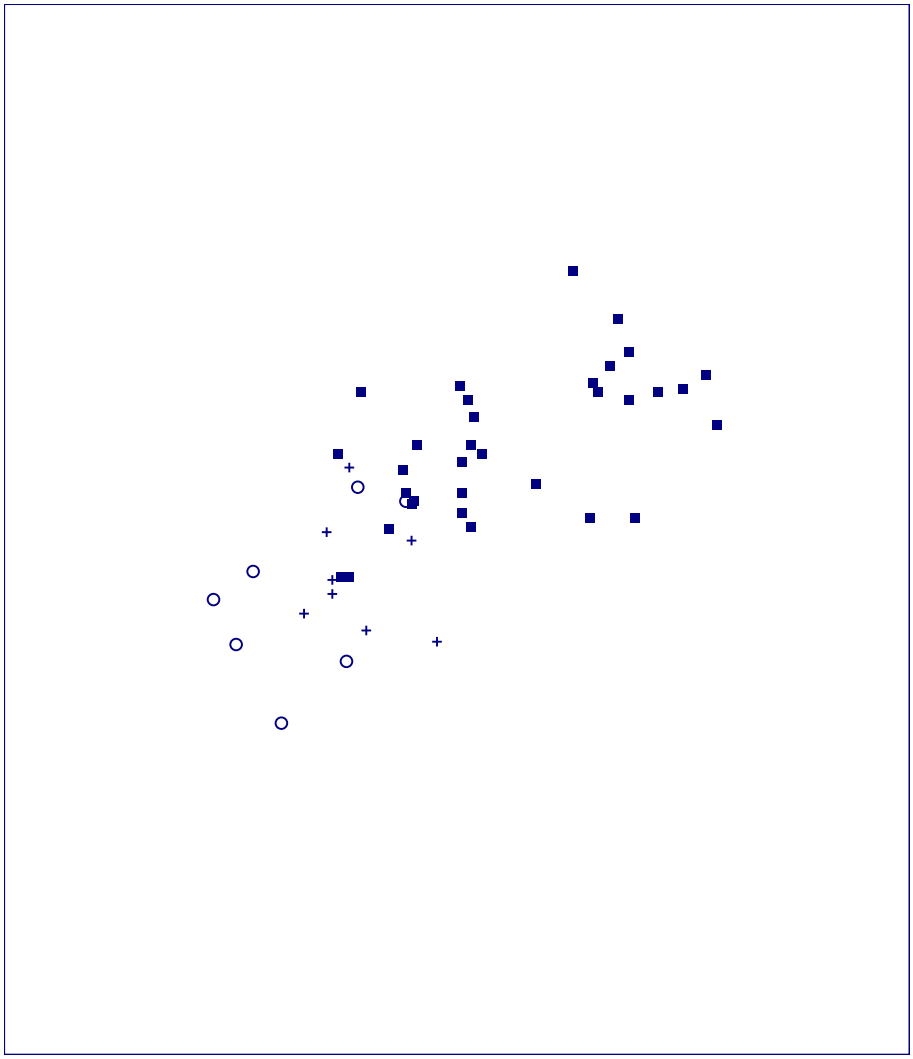
The initial investigation and development of the system was carried out using the ^{13}C data set. This data was very suitable for a preliminary study as the signal-to-noise ratio was relatively high and because it had already been ascertained that it was possible to discriminate between the the two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

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


Amplitude
185
170
155
140
125
110
95
80
65
50
35
20
5
-10





Pearson correlation coefficients were used for this analysis since these were more appropri-

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)***

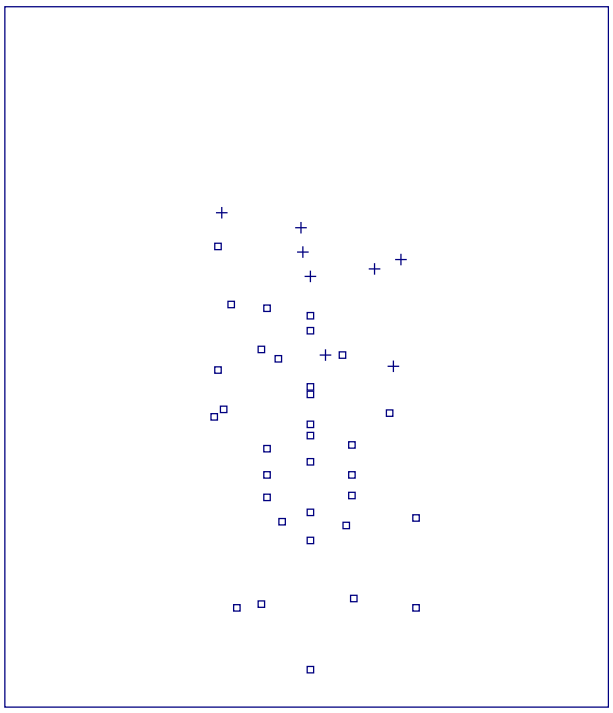


Table 5.4. Classification results using linear discriminant analysis employing wavelet coefficients when the vegetarians were included.

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

5 Conclusions

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

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 vivr***

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

as input and produces as output a classification for the signa

Stage 2 Classification

- The classification for the spectrum, together with the proba

[Abbott, 1994

[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,

[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***

[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.

[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***