OPTIMISED SPECTRAL PROCESSING AND LINESHAPE ANALYSIS IN 2-DIMENSIONAL J-RESOLVED NMR SPECTROSCOPY BASED METABOLOMICS

by

HELEN MICHELLE PARSONS

A thesis submitted to

The University of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY

School of Biosciences

The University of Birmingham

August 2009

Abstract

NMR spectroscopy is a primary analytical approach of metabolomics. Although 1D ^{1}H NMR spectroscopy is versatile, highly reproducible and widely used, analysis of complex biological samples yields congested spectra with many overlapping signals. This makes metabolite identification and quantification challenging. ^{1}H J-resolved (JRES) experiments spreads this high signal density into a second dimension, simplifying the spectral analysis. This thesis analyses the approaches and suitability of JRES spectroscopy to analyse metabolomics data.

Firstly, the robustness of the JRES experiment is investigated. Using spectral relative standard deviation, benchmarks of spectral robustness can be compared between disparate processing techniques, sample types and analytical platforms. JRES spectra were found to be suitable for metabolomics experiments. Secondly, the application of standard metabolomic analysis methods to JRES spectra was examined. Using principal component analysis, the classification accuracy of 1D ^{1}H and JRES spectra were investigated using several data sets. Alongside, three scaling methods were also evaluated. It was found that 2D JRES spectra and the glog transformation could produce 100% classification accuracy. Finally, spectral deconvolution of 2D JRES spectra from line-shape fitting was investigated Here, the mathematical functions describing the JRES line-shape, under several different processing conditions, are derived and used to create a semi-automated metabolite identification and quantification algorithm. Furthermore, possible quantitation errors arising from using JRES spectra are investigated, evaluating effects such as the overlapping of dispersive tails of nearby signals.

In conclusion, the JRES experiment is a suitable for use in the field of metabolomics.

Declaration

I confirm that this work is my own and that I have been involved in the design and conduct of these studies, analyses of data and preparation of this thesis. The following aspects of these studies were undertaken as part of collaboration:

All NMR spectra were prepared and collected by various collaborators as noted in the respective chapters of this thesis. Collaborators include Drs Adam Hines; Huifeng Wu; Stefano Tiziani and Christian Ludwig at the University of Birmingham and Dr Drew Ekman at the U.S. Environmental Protection Agency. Spectral processing, however, was performed primarily by myself unless otherwise stated.

The concept of the extended glog transformation belongs to Drs Christian Ludwig and Ulrich Günther, however, the refinement, testing and applications are my original work.

The preparation and acquisition of all spectra, along with the analysis of the 1D and pJRES spectra found in chapter 4 was performed by Drs Stefano Tiziani and Alessa Lodi. I, however, performed the processing and analysis of the intact 2D JRES spectra.

The measurement of metabolite concentrations by NMR Suite was performed by Dr Mark Viant.

Acknowledgements

First of all, I would like to thank my friends and family - without your support and encouragement I would have never have finished this work. I would also like to thank Dafyd, as you always helped me see the problems in their proper perspective.

Thanks also to my supervisors, Mark, Theo and Kevin as the direction, support and enthusiasm you provided were essential for the generation of this work. I would also like to thank NERC and the EPSRC for funding this research. Many thanks to everyone who has provided me with data throughout the course of this research, especially Adam, Huifeng and Stefano. Also, I would like to acknowledge the thoughtful comments, answered questions and discussion that everyone at the CSB, BLISS lab and HWB gave me, as this was immensely helpful over the last four years. I would especially like to thank Dov, Patsy and John in particular.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and background

1.1 Introduction to metabolomics

Metabolomics is the study of the products of metabolism found in biological samples, and which include many low-molecular-weight compounds such as lipids, sugars and amino acids. These chemicals are often referred to as 'metabolites'. The number and type of metabolites detected in an experiment is greatly dependent upon the source of the sample. This is because the number of different metabolites varies greatly between organisms and even between tissues. For example; for humans, estimates of the number of metabolites present range from around 2,000 up to 20,000 [58]; the fungus known as bakers yeast, *Saccharomyces cerevisiae*, has around 584 [11]; whilst the plant kingdom has an estimated 200,000 primary and secondary metabolites [7]. The analysis is further complicated by the fact that not all metabolites are present in all sample types, which are many and varied: tissue samples [22]; biofluids [39] such as blood plasma or urine [72]; samples of bacteria and cell extracts [65]. Finally, there is also the wide variety of analytical methods used; such as chromatography, nuclear magnetic resonance (NMR) spectroscopy and mass spectroscopy (among many others) [6, 38]; each detects a further subset of metabolites present.

One of the most popular methods in metabolomics is that of NMR spectroscopy, due to the wealth of information that may be extracted from a spectrum obtained in a relatively quick and simple manner [6, 71]. The high-throughput nature of many experiments, such as the very common one-dimensional (1D) proton (^1H) experiment, also allows large data sets to be quickly and rigorously analysed [77]. The non-destructive nature of the experiments also allows subsequent analysis if necessary by other techniques [10]. Unfortunately, NMR suffers from several drawbacks such as spectral congestion [71], overlapping of small resonances by intense peaks [62] and difficulties in comprehensive metabolite identification (see for example, Hines et al. [22], where most indicators of hypoxic stress are identified, but a highly significant singlet is only identified by chemical shift value).

Fortunately, not all NMR experiments are as susceptible to these drawbacks. J-resolved (JRES) NMR, for example, is an experiment type suitable for analysing small molecules - such as metabolites - where the resonances are spread across two acquisition dimensions, reducing spectral congestion. For this reason, JRES spectra are becoming more prevalent in metabolomics [71, 79, 75]. However, the full surface spectrum is not fully exploited, instead only the projections of the spectrum are calculated - pJRES spectra - which are then used in the same manner as the 1D ^1H spectra [22, 12]. However, although this diminishes the problem of spectral congestion, working with pJRES spectra eliminates the data present in the second dimension (i.e. the spin-spin coupling), discarding information that could be used to help identify metabolites or contribute to statistical analysis.

Many metabolomic experiments, such as NMR spectroscopy and mass spectrometry (MS), are designed to measure *all* metabolites present in a sample simultaneously. Whilst this ensures that any important information is acquired during a single experiment, it can be difficult to analyse and interpret the data in meaningful and useful ways. Broadly, there are two main approaches used for data analysis: profiling and fingerprinting [17]. Fingerprinting is when the whole of the acquired data (such as an entire NMR spectrum) is used to classify samples into one of the expected experimental groups. For example, to

determine if a sample taken from a patient belongs to a diseased or healthy classification or to determine if these classes are significantly different. In this method of investigation, multivariate statistical data mining techniques such as principal component analysis (see chapter 1.4 for details) are used to analyse the data [17, 32, 37, 69, 71]. The profiling approach uses targeted feature extraction to identify the most relevant parts of the data, potentially reducing the dimensionality of the analysis significantly. Most often, the "features" of interest are the classification and quantification of metabolites present in the samples[10, 78], but may also include identifying signals in noisy data [29].

The use of metabolomics is widespread throughout many fields such as plant biology [7, 77]; medicine[13, 32] and microbiology [11, 41]. In this thesis, however, the main emphasis and source of data is from environmental metabolomics studies [22, 57, 67, 74]. Environmental metabolomics can be defined as "the application of metabolomics to the investigation of both free-living organisms obtained directly from the natural environment (whether studied in that environment or transferred to a laboratory for further experimentation) and of organisms reared under laboratory conditions (whether studied in the laboratory or transferred to the environment for further experimentation), where any laboratory experiments specifically serve to mimic scenarios encountered in the natural environment" [45].

1.2 Aims and objectives

The aims of this thesis can be broadly stated as the evaluation and application of the use of NMR JRES experiment (J-resolved: see section 1.3.3 for more details) to the field of metabolomics. Specifically, this can be broken into three smaller objectives:

1. Spectral robustness. Are 2D JRES spectra as robust as their 1D counterparts? If not, are they able to discriminate effectively between experimental classes and accurately represent biological data?

2. Evaluation of JRES spectra to fingerprinting approaches. Whilst pJRES have been widely used in multivariate analysis such as principal component analysis, are intact 2D spectra also suitable for usage?

3. Ease of feature extraction. Whilst many attempts have been made to automatically extract metabolite identities and quantities, JRES spectra have not been investigated in this manner.

1.2.1 Thesis outline

Chapter 1 is an introduction to the project, which includes sections detailing the relevant background, such as section 1.3, which describes the acquisition, processing and a brief discussion of the mechanisms of NMR spectroscopy and section 1.4, which gives a discussion of the multivariate and mathematical tools used herein.

Chapter 2 deals with the spectral analysis of NMR spectra and focuses upon objective 1. Here, section 2.1 outlines the principle of spectral relative standard deviation (RSD), a tool that can be used to compare spectra as dissimilar as pJRES and those acquired by mass spectrometry. Examples of RSD are discussed in section 2.2, including examples of the spectral RSD of pJRES and JRES spectra.

Chapter 3 contains a discussion on the use of variance scaling metabolomics data, to improve the classification accuracy of the commonly used statistical techniques such as principle component analysis (PCA). Common variance scaling techniques are then compared and evaluated on a range of data types in section 3.3.

Chapter 4 briefly describes the application of 2D JRES spectra to the field of metabolomics and examines optimal processing methods. Chapter 5 then discuses the analytical line-shapes used to describe the peaks in each of the 2D JRES spectra at each stage of the processing procedure.

Chapter 6 uses the results of chapters 4 and 5 to examine the potential application of 2D JRES spectra to produce quantitative metabolomics data. Here, several chemically defined samples are examined along with a biological data set to establish both the proof of principle of the method and compare the results with those generated using 1D spectra.

Chapter 7 contains the summary and conclusions of this thesis, along with descriptions of the need for further work that has been identified during the investigations.

The majority of this work has been taken from publications generated through the course of the research. As such, significant input from other authors such as Mark Viant has been made. This input includes editing of manuscripts, discussion of ideas and the collection of experimental data. However the ideas, research and analysis described in this thesis are the authors own contributions, unless otherwise specifically noted.

1.3 NMR spectroscopy

As the focus of this thesis is upon the analysis of metabolomics data acquired by NMR spectroscopy, this chapter is intended to give an introduction to the process. The exploitation of the phenomena of NMR is sophisticated, complicated and immensely useful in a number of scientific and medical disciplines. Whilst NMR spectroscopy is not as sensitive as other spectral experiments such as ultraviolet and infrared, it has the advantage of being able to collect a range of data about a large number of metabolites simultaneously in a short period of time [7, 80]. NMR analysis is also non-destructive and does not require any selection of analysis conditions prior to the experiment [7]. These advantages make NMR an ideal tool for metabolomic studies that necessitate the analysis of large numbers of samples.

This section provides a brief overview of the collection, analysis and structure of 1D ^{1}H NMR and 2D J-resolved (JRES) experiments; focusing upon concepts discussed later in

this thesis. The topics discussed include: the theory of NMR, properties and processing of the JRES experiment; general spectral post-processing methods and sources of spectral variation.

1.3.1 Theory of NMR

The description given in this section is based upon the classical theory of NMR, largely omitting the more complicated formal quantum description. The so-called "vector model", whilst not a true picture of the process of NMR, provides a good approximation to the ideas and fundamental principles essential to understanding the analysis [4, 15, 23].

Nuclear spin and magnetism

A fundamental property of atomic nuclei that allows NMR experiments to be performed is that of *spin*. This intrinsic angular momentum is a quantum property where the quantized magnitude is calculated as $\hbar\sqrt{I(I+1)}$. $\hbar$ represents Planck's constant divided by 2π and I is the spin number associated with the nucleus. The spin number is dependent upon the number of atomic particles contained within the nucleus and may only take limited values, usually $\frac{n}{2}$ where n is an integer. In this work, all NMR spectroscopy conducted uses protons (^1H) for which $I = \frac{1}{2}$ [70].

Along with the magnitude, the direction of the spin is also quantized, with each nuclei having exactly $2I + 1$ orientations upon an arbitrary z-axis. For example, protons have only two orientations which are shown diagrammatically in figure 1.1. Transitions between the different orientations may occur, but only using integer steps i.e. $1 \rightarrow 2$ or $-\frac{1}{2} \rightarrow -\frac{3}{2}$ etc. The angular momentum $\boldsymbol{I}$, being a vector quantity, may be resolved against an arbitrary axis, z:

$$I_z = m\hbar \tag{1.1}$$

where m is the magnetic quantum number. m is closely related to the spin number, I, and takes the values $-I, -I+1, \ldots 0, \ldots I-1, I$. Each of these values is associated with

the orientation of the nucleus against the arbitrary axis (see figure 1.1).

A nucleus with spin can also be treated as a rotating magnet (or a *dipole*) and hence must also have a magnetic moment associated with the angular momentum. The magnetic moment can be calculated as:

$$\boldsymbol{\mu} = \gamma \boldsymbol{I} \tag{1.2}$$

where γ is called the gyromagnetic ratio which is specific for each species of nuclei. For protons $\gamma = 26.75 \times 10^7 T^{-1} s^{-1}$ [23]. A positive γ indicates that the magnetic moment is parallel to the spin, whilst negative γ indicates that they are anti-parallel.

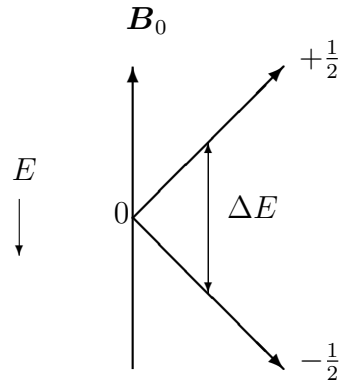


Figure 1.1: Figure illustrating the change in energy, ΔE, associated with the change between the two nuclei orientations of protons, $m = \pm\frac{1}{2}$ in the external magnetic field $\boldsymbol{B}_0$. The negative state is defined to have higher energy.

In the presence of a extraneous magnetic field, $\boldsymbol{B}_0$ (aligned with the arbitrary z-axis, see figure 1.1), the $2I+1$ different orientations of the nucleus are excited from the equilibrium positions and begin to precess about the axis of the applied magnetic field. The frequency of rotation of each magnetic moment is again unique to each species of nucleus and is called the *Larmor frequency* [15]:

$$\omega_0 = -\gamma B_0 \tag{1.3}$$

where $B_0 = |\boldsymbol{B}_0|$. The application of the magnetic field also alters the distribution of energy within the system. Since energy may be calculated as [23]:

$$E = -\boldsymbol{\mu} \cdot \boldsymbol{B}_0 \tag{1.4}$$

quantized energy levels are present within the system, dependent upon the orientation of the spin of the nuclei (see equation (1.2)). By "flipping" from one energy level to another, the nuclei emits the change of energy (ΔE) of a specific frequency, ν. This phenomena is called the *resonance frequency* and can be written as [23]:

$$\Delta E = h\nu \tag{1.5}$$

where h is Planck's constant. The resonance condition is of vital importance to NMR spectroscopy, as the energy emitted, ΔE, is the processed that is measured [15].

To calculate the resonance frequency of a nucleus, firstly consider the energy of a nucleus (equation 1.4) along the z-axis aligned with the magnetic field:

$$E = -\mu_z B_0 \tag{1.6}$$

Using the z-component of equations (1.1) and (1.2), equation (1.6) then becomes:

$$E = -m\hbar\gamma B_0 \tag{1.7}$$

It is also known that energy transitions may only occur in unit steps (i.e. $\Delta m = \pm 1$), which then gives :

$$|\Delta E| = \hbar\gamma B_0 \tag{1.8}$$

Combining equations (1.5) and (1.8) then yields:

$$\nu = \frac{\gamma B_0}{2\pi} \tag{1.9}$$

Although, in very simple terms, NMR spectra are formed by the observation of this energy when the atomic nuclei "flip" between the energy states arising from the application of the magnetic field, the phenomenon has many other complications and subtleties. Firstly, the atomic nuclei in a sample are usually part of a "lattice" - that is, coupled together to create the structure of a solid or liquid, referred to as the spin-lattice coupling. Thus the spinning magnetic moments of each nucleus (called the spin system) and the lattice of the sample are closely related and evolve towards a common thermal equilibrium. When the constant magnetic field $\boldsymbol{B}_0$ is applied to the system in equilibrium as described above (figure 1.1), only the longitudinal component of the magnetic moment is non-zero i.e.; $M_z = M_0$ and $M_x = M_y = 0$. If the system is not in equilibrium, the magnetic moments will evolve to the equilibrium conditions through an exchange of energy between the spin system and the lattice structure. It is assumed that the rate of progression of the magnetic moment towards the equilibrium state, or *relaxation*, of the nuclei is exponential in each dimension: T_1 is used to represent the longitudinal component along the z-axis and T_2 the transverse component along the x and y axes. This can be written as:

$$\frac{d}{dt}M_z = -\frac{1}{T_1}(M_z - M_0) \tag{1.10a}$$

$$\frac{d}{dt}M_{x,y} = -\frac{1}{T_2}M_{x,y} \tag{1.10b}$$

which forms the basis of the Bloch equations in a stationary reference frame:

$$\frac{d}{dt}\mathbf{M} = \gamma\mathbf{M} \wedge \boldsymbol{B} - \frac{1}{T_1}(M_z - M_0)\boldsymbol{k} - \frac{1}{T_2}(M_x\boldsymbol{i} + M_y\boldsymbol{j}) \tag{1.11}$$

where $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ are unit vectors along the x, y, z axis respectively. More details on the Bloch equations, including the modifications needed to describe a rotating magnetic field can

be found in many NMR textbooks, such as Goldman [15].

Historically, a continuous magnetic field was applied to excite the nuclei to the state of higher energy ($m_s = -\frac{1}{2}$). Known as the continuous wave (CW) method, this simply involves applying a magnetic field for several minutes whilst the energy frequency range is changed by either altering the magnetic field or frequency of the radio transmitter (the mechanics of the spectrometer are discussed later in greater detail).

Alternatively, the more modern Fourier Transform (FT) method can be used, which is the method used exclusively in this thesis. When using the FT method, the magnetic field is applied in the form of a radio frequency (rf) pulse consisting of the entire frequency range of interest, lasting only a few microseconds so that the pulse duration, t_p, is far smaller than either relaxation parameter. This generates an oscillating magnetic field, $\mathbf{B}_1$, along the x-axis which alters the resultant direction of the total magnetic field by an angle of $\theta = \gamma B_1 t_p$. Commonly, the time of the pulse t_p is chosen such that the pulse angle of the resultant field is a predetermined value, such as 90° (called a $\frac{\pi}{2}$ pulse) for example. Once the resultant magnetic field has moved from its orientation along the z axis by the rf pulse, the magnetic moment then precesses around the z axis in the xy plane until the system relaxes into its equilibrium state. This can be stated as [15]:

$$M_x = M_0 \sin\theta \cos\omega_0 t \exp\left(\frac{-t}{T_2}\right) \tag{1.12a}$$

$$M_y = M_0 \sin\theta \sin\omega_0 t \exp\left(\frac{-t}{T_2}\right) \tag{1.12b}$$

$$M_z = M_0 \left\{1 - (1 - \cos\theta) \exp\left(\frac{-t}{T_1}\right)\right\} \tag{1.12c}$$

A receiver coil parallel to the x axis will then detect an induced voltage V proportional to M_x. However, this signal oscillates as the magnetic moment precesses along the positive x axis (positive x signal) towards the y axis (zero x signal) then along the negative x axis (negative x signal) and so on. Since the rf. pulse produces resonance effects at all

frequencies, a superposition of the voltages from each of the nuclei present is detected. A FT is then necessary to convert the detected voltage from a series of cosine functions in the *time domain* called the "free induction decay" (FID) to an easily recognisable spectrum of resonances in the *frequency domain*, as shown in figure 1.2.

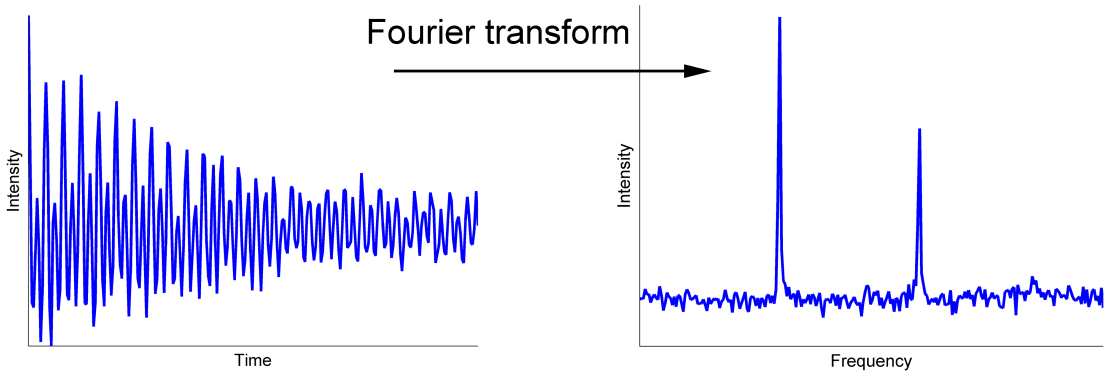


Figure 1.2: The Fourier transform takes the acquired signal from the time to the frequency domain

Chemical shift

Clearly, the resonance frequency of a nucleus is dependent upon the magnetic field that is applied (equation (1.9), page 9); however, other factors such as the chemical bonding of the nuclei also alter the frequency to a lesser extent, due to a phenomenon known as *shielding*. This *chemical shift* is another vital property of NMR spectroscopy, as detecting these (relatively) small changes in resonance frequency allows unique chemical structures to be identified and provides an NMR 'fingerprint' for each metabolite. Measuring absolute frequency for each separate nucleus in a sample can be cumbersome as the change in frequency can be several orders of magnitude smaller than the initial value. For example, subject to a 9.4T magnetic field, protons resonate at approximately 400MHz, yet a typical ^{1}H NMR spectrum spans around 4kHz about this initial (operating) frequency [23]. Hence NMR spectra are typically reported as a relative frequency, using an internal standard - a known compound added to the sample. Internal standards are chosen to as they resonate at high frequencies (due to their highly shielded nuclei) and chosen

so they create a single peak far from resonances of interest. This then makes the measurement of frequency difference between any peak and the internal standard simple and peak interference unlikely. Possible internal standards include the compounds sodium 3-trimethylsilyl-2,2,3,3-d_4-propionate (TMSP); tetramethylsilane (TMS) and sodium 2,2-methyl-2-silapentane-5-sulfonate (DSS). The internal standard used throughout this thesis is TMSP.

By measuring the difference in the frequencies of the resonance peak of interest, ν_p, and the internal standard ν_S (both measured in Hz), the chemical shift scale δ can be calculated as:

$$\delta = \frac{\nu_p - \nu_S}{\nu_0} \tag{1.13}$$

where ν_0 is the operating frequency of the spectrometer, usually measured in MHz. Clearly, δ is a dimensionless quantity, but it is commonly expressed as fractions of the applied field in parts per million (ppm). By convention, the chemical shift scale is reported such that resonances downfield that yield high frequency and chemical shift values appear before low frequency and chemical shift values upfield, as shown in figure 1.3. The typical range of frequencies studied in ^1H NMR form the chemical shift range of 0-10ppm.

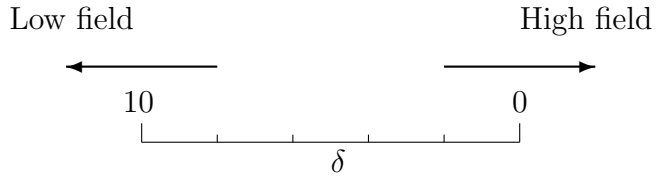


Figure 1.3: Illustration of the chemical shift range for ^1H NMR. δ is reported such that resonances downfield that yield high frequency and high chemical shift values appear before low frequency and chemical shift values upfield.

By reporting spectra in terms of chemical shift, the dependency upon the magnetic field strength is removed, allowing spectral comparison between different spectrometers. Chemical shift also has the added advantage of being a more simple number to report than the specific frequency in Hertz. This is since the frequency range of approximately

400MHz ±4kHz [23] requires a large number of significant figures to report accurately small peak shifts.

Intensity of NMR resonances

For ^1H NMR spectroscopy, the integral of the resonance is proportional to the number of protons being detected, provided that all nuclei are fully relaxed i.e, that all nuclei return to their equilibrium state between the pulses in a FT-NMR experiment. This condition is mostly fulfilled in proton spectroscopy, but is not the case when other nuclei - such as ^{13}C - are used. Note also that due to this proportional nature of the peak intensity, the intensity axis of NMR spectra is also usually expressed as a dimensionless quantity.

Robustness of the chemical shift

As stated above, there are many factors that affect the resonance frequency. Whilst these changes allow the nuclei in different molecular structures to be distinguished and identified, there are also factors that that can change the chemical shift of the same molecular structure between experiments. These factors include [80]:

- The inductive effect, where nuclei in a region of high electron density exhibit a change in resonance frequency. These nuclei are *shielded* by the electrons, as a secondary magnetic field is created by the circulating charges, changing the overall applied magnetic field.

- Bond interactions. Similar to the inductive effect, bonds between nuclei are regions of high electron density, creating small changes in the magnetic field. This effect is discussed in detail in the section "Spin-spin interaction" below.

- The temperature of the sample also affects the chemical shifts of some peaks. For example, at higher temperatures there is reduction in the degree of hydrogen bonding in OH and NH molecules.

- Solvents also can affect the resonance frequency of the nuclei. Added to the sample as part of experimental preparation to ready the sample for use, the choice of solvent can vastly change the created NMR spectrum. For example, benzene weakly solvates in areas of low electron density and so can obscure signals of solute nuclei nearby. Also, changing between a solvent such as CCL_4 to a polar solvent such as acetone can change the chemical shift of a resonance up to 0.3ppm.

- Changing pH levels can also alter the electron density of certain susceptible structures.

Combinations of these effects can change the values of the resonance frequencies of the different nuclei in a molecule in a non-linear manner, even within a single molecular structure. For example, figure 1.4 shows the pJRES spectrum (a type of ^1H NMR experiment. See section 1.3.3, page 18, for more details) of pure histidine at three different pH values. It can be seen that the peak at approximately 8ppm is highly susceptible to changes in pH, whilst other peaks are largely invariant to the altered environmental conditions.

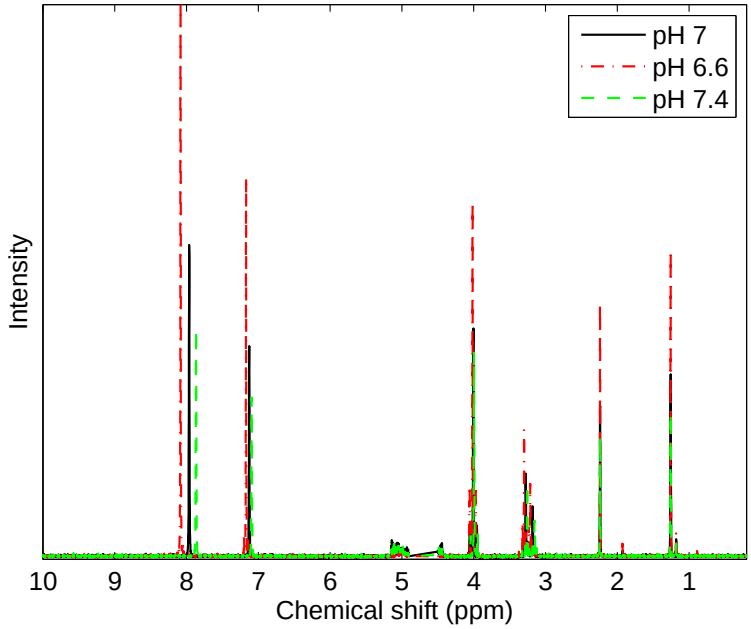


Figure 1.4: pJRES spectrum of histidine acquired at three pH values. It can be seen that the peak at approximately 8ppm is highly susceptible to changes in pH, whilst other peaks are quite robust

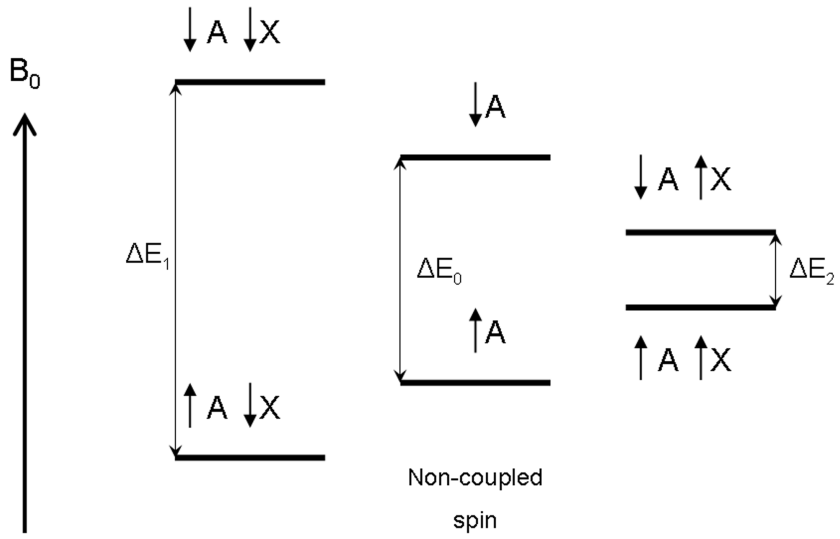


Figure 1.5: Spin-spin coupling interactions of an AX (weakly coupled) system under the applied magnetic field $\boldsymbol{B_0}$. For the single nuclei, A, the energy emitted is ΔE_0, however when coupled with nuclei X, the alignment of the nuclei alters the emitted energy $\Delta E_1 \neq \Delta E_2$. Energy level separation is not drawn to scale.

Spin-spin interaction

Previously, whilst it has been discussed that resonances arise from each proton (or nucleus type being excited), it has been implied that a single resonance is generated by each nucleus. In fact, this is a simplification of the experimental processes and completely ignores the phenomena of *spin-spin interaction*.

The magnetic dipoles of each nucleus in a single molecule are sufficiently close to one another to affect each other through the electrons of the bonds between them. This results in an effect known as *splitting* and the spins are said to be *coupled*. Depending upon alignment of the nuclear spin, the magnetic field produced (from the spinning charge) augments or weakens the total magnetic force acting upon the coupled nuclei. This, in turn changes the energy associated with, or *energy level* of, that alignment which then alters the energy emitted during the transition between states.

As shown in figure 1.5, when a nucleus A (of any spin) is coupled to a nucleus X (of any spin), two distinct energy transitions may take place, depending upon the orientation of the X nucleus. Note also that only single spin transitions are allowed. Since each of these transitions emits energy of a different frequency, *two resonance peaks*, called a doublet, are detected for the single nucleus spin A. These two peaks appear centered at chemical spin of δ_0 ppm, separated by a distance J_{AX}. J_{AX} is known as the *spin-spin coupling constant* between the nuclei A and X and is reported in Hertz. More generically, the spin-spin coupling constant is reported as J, and may take either positive or negative values. Whilst the sign of J has no effect upon the separation of the peaks, it indicates the predominate arrangement of the spins: $J > 0$ shows that the anti-parallel (i.e. the higher energy state, aligned against the magnetic field) is the dominant spin arrangement, whilst $J < 0$ denotes that the lower energy level parallel with the magnetic field is more common [23].

Again, figure 1.5 illustrates only a simple example of spin-spin coupling. Adding more spins to the system, splits the peaks into further complex multiple peak structures - called multiplets - such as triplets (a resonance split into three peaks) and quartets (four resultant peaks) with specific intensity patterns. Further splitting patterns are discussed in Williams and Flemming [80]. Another complication comes from the mechanism known as *strong coupling*. Strong coupling arises from coupled spins whose nuclei are relatively close ($|\delta\nu| \approx |J|$). Here, the intensity patterns of the nuclei in the system begin to overlap and interfere with each other, distorting the multiplet pattern. A further discussion of strong coupling can be found in Hore [23].

1.3.2 Signal detection

NMR spectrometers are complicated instruments capable of conducting many different types of experiment. Put simply, however, the vital components of a spectrometer consists of a strong magnetic field; a sample probe; a radio frequency (r.f.) transmitter; a r.f.

receiver and a computer to digitize, record and analyse the signals. A schematic diagram is shown in figure 1.6 [23].

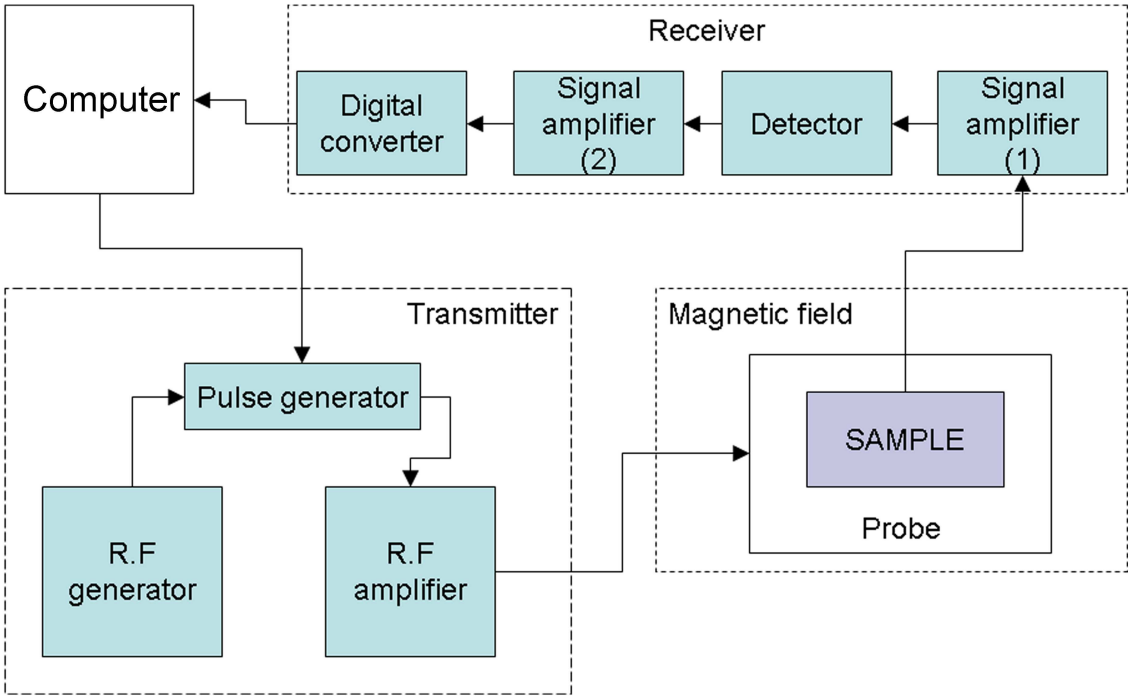


Figure 1.6: A schematic diagram of the main component of a NMR spectrometer. Here, R.F stands for radio frequency.

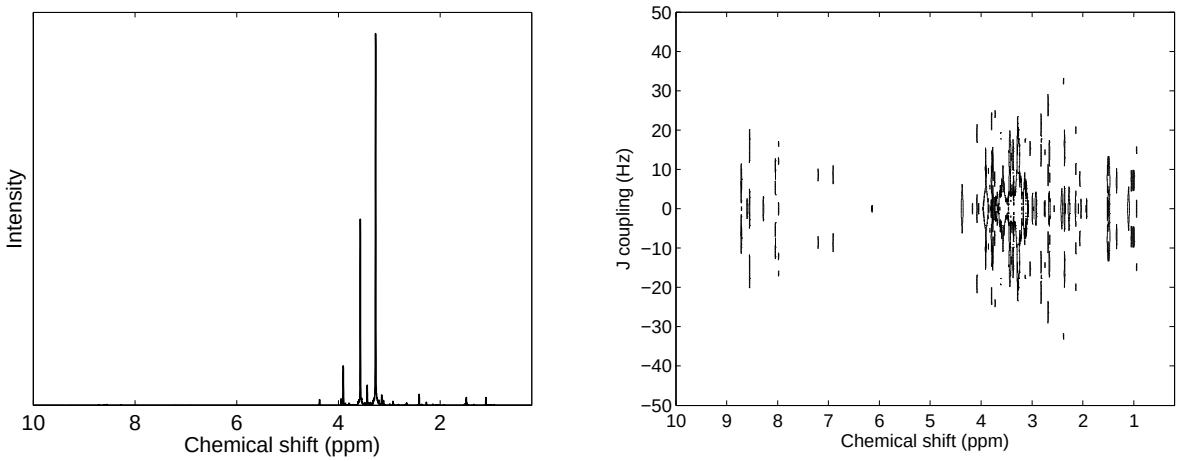
Of particular interest to this work is the mechanism of the detector, labeled as the 'receiver' in figure 1.6. Here, the signal acquired from the probe is amplified, then detected using a phase sensitive detector. This signal is amplified again, then split into two components which are 90^o phase shifted (referred to as "real" and "imaginary") in a process called *quadrature detection*. This process is important as it alters the signal-to-noise ratio (SNR) of the final spectrum [14]. It is also important to note due to this method, the primary source of error - the "noise" - present in an NMR spectrum originates from the thermal noise generated through the electrical transmission of the signal. It is also known that prior to digitisation, the noise is uncorrelated and Gaussian distributed, but that the signal processing subtly changes this as described in Grage and Akke [18].

1.3.3 JRES NMR

The FT-NMR experiment described in section 1.3.1 is only a fraction of the different experiment types possible using NMR spectroscopy. These experiments are termed one-dimensional (1D) spectroscopy, since only one frequency axis is used to record the data. By using multiple r.f. pulses, two-dimensional (and three or even four-dimensional) experiments can be conducted. These multi-dimensional experiments form multiple orthogonal frequency axes along with an intensity axis. Depending upon the pulse sequence used, the two frequency axes display different information. For example, a COSY (**CO**rrelated **S**pectroscop**Y**) spectrum highlights all spin-spin coupled protons [80]. 2D experiments are acquired and processed in a similar manner to their 1D counterparts. That is to say; the initial signal is recorded as a function of two time variables, then multiple Fourier transforms are used to construct the spectrum in the frequency domain.

Whilst many 2D experiments are possible, only one type is discussed in this work: J-resolved (JRES) spectroscopy. This is because many of these higher dimensional experiments require substantially longer acquisition times than 1D methods and so are not appropriate for the high throughput metabolomics studies. JRES spectroscopy has been shown to provide spectra with low peak congestion and high metabolite specificity in a short acquisition time, highlighting the potential uses of the experiment [71, 75].

The two frequency axes in JRES spectra are called the direct (chemical shift) and indirect (spin-spin coupling) axes and an example spectrum is shown in figure 1.7B. Here, the chemical shift axis is analogous to the chemical shift axis of the 1D spectra; reporting the resonance frequencies of each nuclei in the sample. Where JRES experiments differ is that the spin-spin (J) coupling information of the peak splitting is separated into the second, indirect dimension. This reduces the peak congestion that can complicate 1D experiments, whilst still retaining all the necessary information required for analysis.



(a) pJRES spectrum

(b) JRES spectrum

Figure 1.7: Examples of a pJRES (part b) and a JRES (part a) spectra. Both spectra are of the same spectrum of a marine mussel mussel [47] (see section 3.3.1 for sample and experiment details).

To date, the processing of JRES metabolomics data has comprised of taking a 1D projection of each 2D spectrum, forming a so-called *pJRES* spectrum shown in figure 1.7A; which provides a simple format for multivariate statistical analysis [71]. However, calculation of a 1D projection discards potentially important data, the spin-spin coupling pattern, which could be used to further discriminate between different metabolites within a complex biological sample. In fact, spin-spin coupling measurements could prove of significant benefit to metabolite identification since they are less sensitive to changes in pH than chemical shift values [43].

1.3.4 General processing methods

There are some aspects of spectral processing that are applied to both 1D and JRES NMR spectra, often at the time that the Fourier transform is applied to the acquired signal. This section provides a short introduction to some important methods used in thesis. The applications to JRES spectra specifically are discussed in more detail in section 4.1.

Apodization functions

Improved signal-to-noise ratios are often achieved through the use of one of many different *apodization* (or window) functions. By the very nature of the experiment, the time dependent signal recorded (the Free Induction Decay or FID) is of decaying exponential form since the signal becomes weaker with each oscillation. Apodization functions are applied to the entire FID and exploit the data format by weighting the initial, strong, signal and dampening the latter, weak signal and noise. Typically, for 1D experiments exponential functions of the form:

$$W(t) = \exp(-Lt) \tag{1.14}$$

are used, where L is some chosen line-broadening constant. The result of application of these functions increases the rate of decay of the exponential FID, leading to a broader line-shapes with smaller maximum values after Fourier transform, yet reduced noise. The effects of window functions upon the JRES experiment is discussed in detail in chapter 4.

While window functions that zero the FID at early time points are useful to minimise baseline artifacts, the use of these window functions leads to a lower signal-to-noise ratio in the resulting spectra and may even result in loss of signals with low intensities [43, 47, 73, 79]. The use of a sine-bell function, for example, increases the weight of the early time points of the FID; thereby preventing the signal-to-noise loss induced by the application of the sine window function alone. These functions are more common in processing of 2D spectra [65, 24] and the applications to JRES spectra are discussed in section 4.1.

Zero-filling

One of the limitations of using the Fourier transform to construct an NMR spectrum, is that the Fourier transform assumes that the input signal from the time domain is infinitely periodic (such as a sine function). This can lead to truncation artifacts and line resolution

issues. To mitigate this issue, zeros are appended to the end of the FID before the Fourier transform is performed in a process called *zero-filling*. Whilst this process does not alter the the information present in the resultant frequency spectrum - nor does it improve spectral resolution - it improves the resolution of the FT and hence also improves the appearance of the spectrum and improves the visibility of the peaks to human analysts.

1.3.5 JRES specific processing

JRES specific processing is applied to each spectrum after Fourier transforms (in each dimension) have been applied. They exploit the characteristic structure of the data in order to increase the signal-to-noise ratio, remove noise artifacts and, in conjunction with spectral projection, simplify the spectrum (i.e., to form pJRES spectra). This processing comprises of two steps: *tilting* the spectrum followed by *symmetrisation*. Both transformations are applied to the entire surface spectrum. Figure 1.8 illustrates the effects of these processing steps for a SEM-apodized singlet, doublet and triplet, using a contour map representation of the spectrum. For example, the doublet has a coupling constant of J Hz and a resonance frequency of ν Hz, producing two signals of intensity $A/2$ at coordinates $(w_{11}, w_{21}) = (pJ, \nu + J/2)$ and $(w_{12}, w_{22}) = (-pJ, \nu - J/2)$.

Tilting the spectrum transforms any point (w_1, w_2) in the spectrum to:

$$(w_1, w_2) \rightarrow (w_1', w_2') = (w_1 - w_2, w_2) \tag{1.15}$$

A schematic showing the effects of tilting is shown in figure 1.9, where the effects of the process upon the spectral peaks can be seen.

Symmetrisation is also a spectrum wide transformation, and for each pair of points $(w_x, w_j), (w_y, w_j)$ (where $y = -x$, so the points are symmetrical about the line $x = 0$Hz),

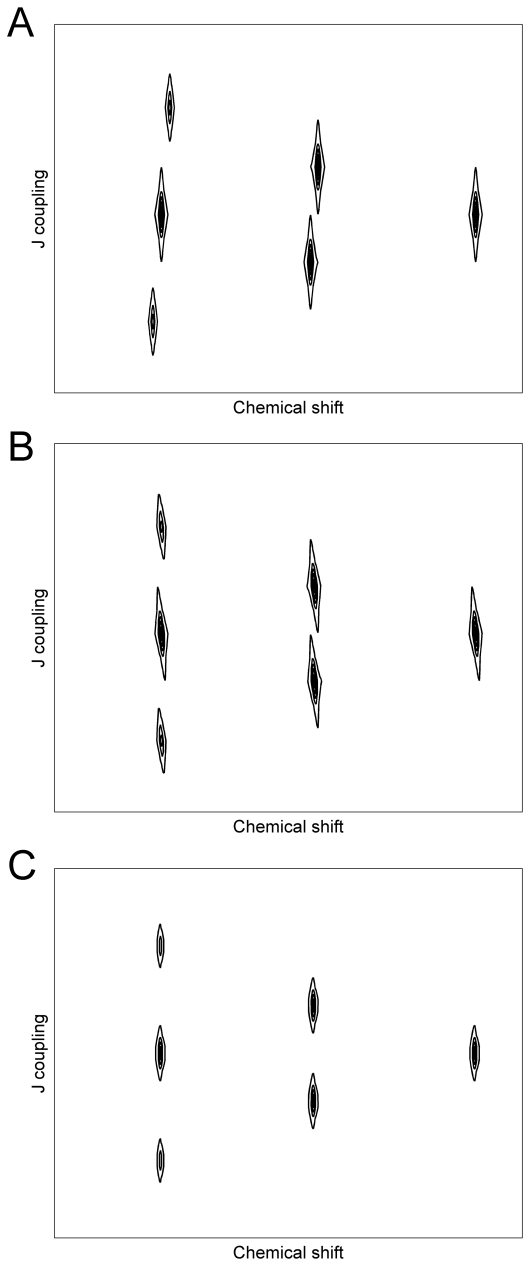


Figure 1.8: Effects of JRES specific spectral processing on simulated singlet, doublet and triplet NMR signals. Contour plots showing: (A) resonances in magnitude mode; (B) application of tilting function to signal in A; and (C) application of symmetrisation function to signal in part B.

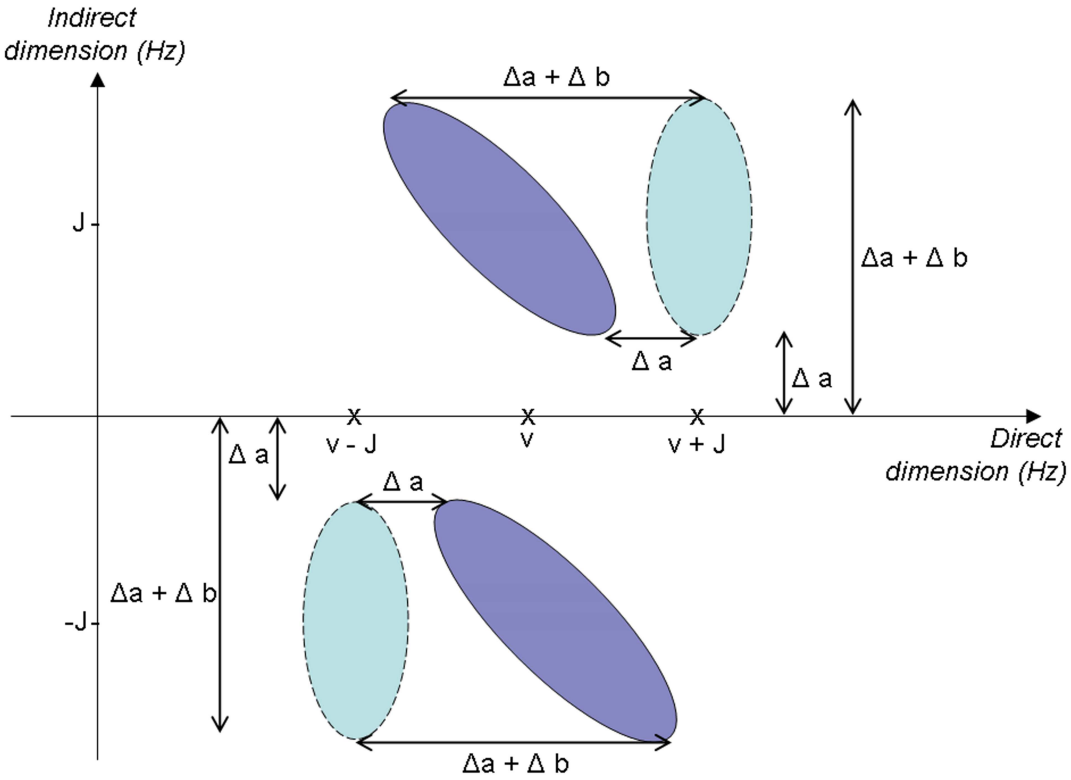


Figure 1.9: Schematic diagram of the tilting transformation. The centres of each peak, prior to the transformation (light blue peaks with the dashed lines), lie at a frequency in the direct dimension of $\nu \pm J$Hz. Tilting then moves the centres of the peaks JHz so all peaks in a given structure (dark blue peaks with solid lines) are aligned at the frequency in the direct dimension of νHz. Points other than the centre are also moved in a similar manner, i.e. each point is translated along the direct dimension by its frequency in the indirect dimension. Since the tilted peaks have been subjected to a shear transformation, the resultant peak shape has been changed from the initial unprocessed spectrum.

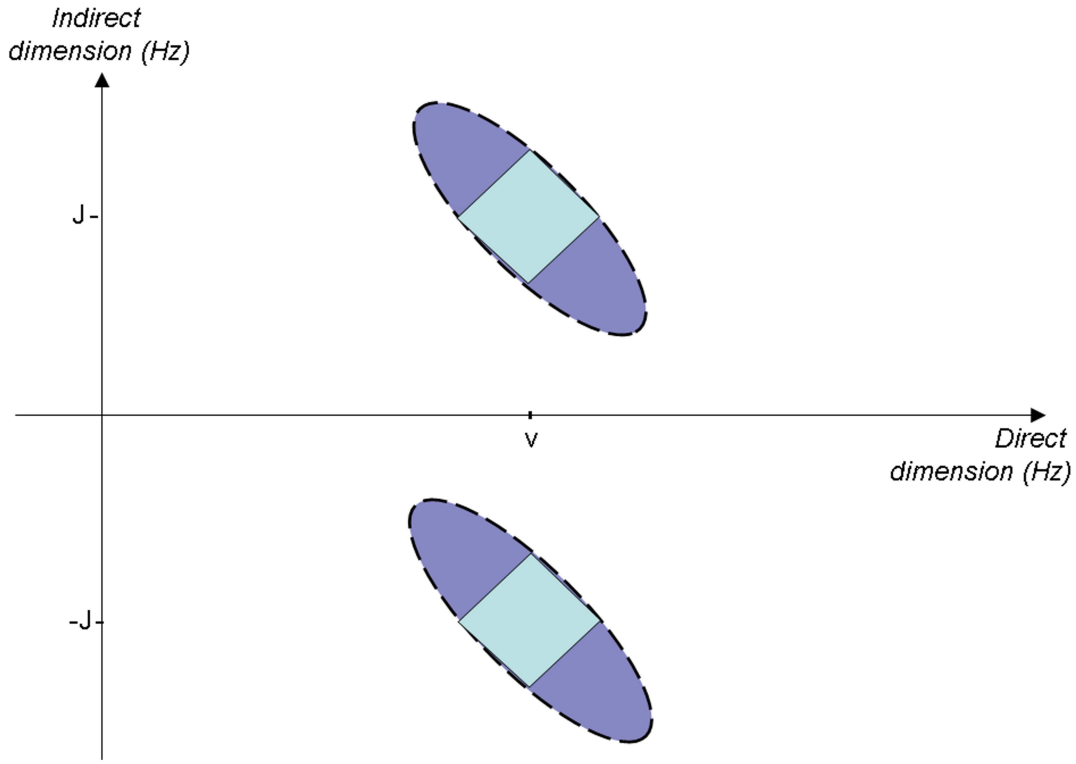


Figure 1.10: Schematic diagram of spectral symmetrisation (or folding). The peaks present after tilting the spectrum (dark blue, dashed lines). After symmetrisation, the peaks (light blue, solid lines) are greatly truncated yet still centred at $(\nu, \pm J)$Hz

the transformation can be written as:

$$(w'_x, w'_j) = (w'_y, w'_j) = \min\left((w_x, w_j), (w_y, w_j)\right) \ \forall j, 0 < x \leq \frac{m}{2} \tag{1.16}$$

where $y = -x$ and m is the width of the indirect dimension in Hertz. Figure 1.10 shows a schematic diagram of the process upon a tilted doublet structure, whilst figure 1.8B illustrates the effects upon several peak structures.

Tilting applies a shear to the spectrum so that the peak maxima in each multiplet now appear at the same resonance frequency in the direct dimension. The spectrum is then symmetrised (figure 1.8C), which forces the signal intensities to become symmetric about the centre line of the spectrum along the indirect dimension. This is achieved by taking the minimum value for each pair of symmetric data points, and effectively discards

artifacts and minimises the spectral noise. This symmetrical, noise reduced spectrum is then easier to analyse as many spectral artifacts and noise spikes are are removed.

1.3.6 Pre-processing methods

Pre-processing is a general term applied to operations performed on series of spectra after they have been acquired from the spectrometer and Fourier transformed into a suitable spectral format; yet before the data are fully analysed (usually by multivariate analysis). These processing techniques are typically applied to improve further analysis of the spectra and to reduce data dimensionality [74, 31, 57]. There are many different pre-processing methods, however most common methods can be grouped into two different categories; dimensionality reduction and spectral refinement.

Dimensionality reduction

A common data pre-processing method used to reduce dimensionality in NMR metabolomics spectra is spectral *binning* [74, 31] (or bucketing [66]). This process segments the spectra into short sections called 'bins' along the chemical shift axis, reducing the amount of data points per spectrum significantly. Typically, the width of the bins are calculated using increments measured in chemical shift [71], but can also be based on the number of data points integrated[65]. Although binning reduces the resolution of the spectrum, this is a useful technique as small, unwanted shifts in chemical shift value is a common problem in NMR experiments, since peaks are susceptible to slight fluctuations in experimental conditions such as temperature and pH (described in section 1.3.1). Binning reduces the number of data points that defines the peak, thereby reducing the impact of these small shifts. This is immensely useful since many analysis tools assume that for each metabolite, peaks occur at identical points in each spectrum.

Other methods to reduce dimensionality involve simply using small sections of the spectra known to contain information of interest [26] and 'clipping' the spectra into consistent

regions, usually the 0.5-10ppm range[74].

Spectral refinement

These pre-processing methods do not alter resolution of the spectra, but are used to reduce specific instances of unwanted variation. There are five main categories:

1. Normalisation. To reduce errors from samples of disparate masses or volumes, all spectra are normalised. There are many different types of normalisation, such as setting the total spectral area to a constant value or setting the area of the internal standard to a specific value.

2. Bin compression. Sometimes the peak shifts of individual metabolites are so great that binning does not align the shifted peaks. In this case, bins can be merged or compressed together, ensuring that each of the shifting peaks are contained in a single bin at a constant chemical shift.

3. Section removal. Most metabolite samples must be combined with some type of solvent (see Lin et al[36] for examples) which have no bearing upon the analysis of the experiment. These peaks can dominate the spectrum due to the sheer quantity of the compound present, biasing results. These peaks are often simply 'cut' from the spectrum and not entered into any further analysis.

4. Spectral alignment. Alignment methods can be used to ensure that peak shifts throughout the spectra are kept to a minimum and so do not affect subsequent analysis[35].

5. Scaling and transforms. NMR spectra do not conform to many assumptions of common statistical tools such as displaying a normal distribution and independence of data points. Various scaling methods and transforms have been used to force the data into a form that can be more easily analysed. This is discussed in further detail in chapter 3 where the application of variance stabilising methods are explained.

1.3.7 Spectral varianace

Variance structure of NMR spectra is an often overlooked, yet vitally important part of data preparation. This is since many commonly used data mining tools such as principal component analysis (PCA), partial least squares discriminant analysis (PLS-DA) and many other tools which rely extensively on variance are used to analyse metabolomic data sets. It is therefore important to assess that the variance structure of the data is appropriate for use with the analysis tool. Unfortunately, NMR spectra violate the basic assumption of homoscedasticity (constant variance) that many mathematical tools assume. This is simply because a point on an NMR spectrum, y can be modeled as

$$y = b + \mu e^{\eta} + \epsilon \tag{1.17}$$

where b is the baseline, μ the true signal value. η and ϵ are both random normally distributed errors with mean 0 and variance 1 [56, 83]. That is to say that the larger the metabolite peak in the spectrum, the larger the variance of that data point between samples. This biases most multivariate analyses towards the most intense peaks in the spectrum, overlooking the smaller absolute variance of less intense metabolites. Therefore, it becomes necessary to modify the variance structure of the data to avoid misleading test results.

Types of variation

Variation between samples can be furthermore classified into one of two types - *technical* and *biological*.

Technical variation is created by the experimental procedure, such as sample preparation and analytical measurement errors; whilst biological variance is the inherent variation between samples created by genetic differences, pathological or environmental factors, etc. The technical variance does not contribute any useful information to discriminate between

different biological sample classes and so, ideally, this variance would not contribute to any data analyses. Data processing methods can be used to affect the structure of the variance of experimental data sets, helping to focus the multivariate analysis onto more biologically relevant information that is more interesting to the experimenter. However, many of these processing methods alter the data so radically that becomes difficult to identify artifacts and other errors.

Biological variation can be further divided into two classes: *interclass* and *intraclass*. Interclass variance is that variation that arises from the different spectral profiles exhibited by the differences between the experimental classes. This variation is usually the main focus of a metabolomics experiment e.g. investigating the differences between healthy and diseased samples. Intraclass variation, however, is the result of biological differences inherent between samples. This can include any biological variation not directly under study, such as age and gender, for example. While it is beneficial to be aware of these differences, most biological variation of any type can be illuminating and of interest to the experimenter.

1.4 Multivariate analysis

Analysing NMR spectra can be time consuming and difficult. Identifying each resonance by hand yields a great deal of information about the sample, but such problems as overlapping peaks and changes in chemical shift can make such a task daunting, even to an experienced analyst. Coupled with the typically large datasets used in metabolomics studies [6], it is clear that automated methods or other analysis methods are needed.

Multivariate analysis is a group of mathematical tools used to simultaneously analyse numerous variables. Whilst using these tools to analyse spectra will not typically give quantitative results, they do allow class differences to be measured. By identifying how inter-class samples differ (i.e. finding the difference between healthy and diseased samples)

this then allows subsequent investigation to be more focused upon known spectral regions of interest.

Types of multivariate analysis methods

Multivariate methods can be split into two different types of analysis: *supervised* and *unsupervised*. Supervised methods use prior knowledge to help analyse data and are widely used where a database of known samples are used to predict the behavior of new samples. Supervised methods can include classification techniques such as as well as regression methods such as partial least squares [47, 76]. Unsupervised methods solely rely on the data input and include the technique *principal component analysis* which is used throughout this thesis. Using unsupervised multivariate techniques ensures that the analysis results have widespread applicability to other tools as well as ensuring fair comparison between different analysis.

1.4.1 Principal component analysis

Principal component analysis (PCA) is one of the main analysis tools used throughout this thesis. It is an unsupervised multivariate technique and so is suitable for use with most NMR metabolomics data investigations. PCA is particularly useful as it reduces a large number of intercorrelated variables to a smaller number of variables, whilst retaining most of the variation [28]. These output variables can be viewed in two ways: as *scores*, or as *loadings*. The scores are simply the input samples transformed into the new coordinate system, whilst the loadings are the correlations of the new variables in terms of the old coordinates. This flexibility of output makes PCA very popular in NMR metabolomics [22, 37, 47, 73].

PCA is also a useful tool as it makes no prior assumptions about the data - solution groupings are detected without any bias added by the experimenter. For example, Hines *et al* [22] conducted a PCA of a group of mussels in two classes (hypoxic and control).

However, the analysis revealed an additional two groupings (later to be identified as gender) which became unexpectedly important results - only identified due to to the variable-directed nature of PCA [76].

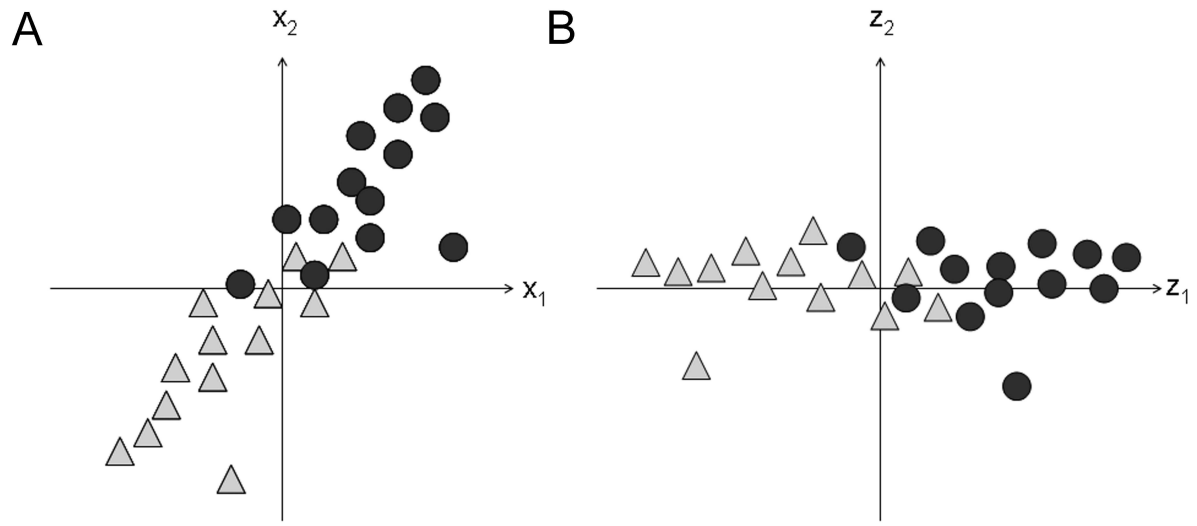


Figure 1.11: Illustration of the scores plot of PCA on a simple 2D example. The initial variables x_1 and x_2 (part A) are transformed such that the resultant principal components z_1 and z_2 (part B) express the maximal components of the data variance.

Other advantages of PCA are not so obvious, but are also useful for the exploratory analysis of NMR metabolomics data. PCA produces an orthogonal set of unique components, from which it is a simple task to determine the linear combination of the initial variables used to produce each output principal components. PCA also uses very few assumptions as there is no underlying statistical model [76] and relies purely upon the variance structure of the data for the analysis [28]. As shown in figure 1.11, the principal components created by the technique illuminate the variables that explain the maximal variance structure of the data, helping quickly to identify any trends or groupings.

Other feature extraction techniques such as factor analysis (FA) and independent component analysis (ICA) use properties such as the covariance structure or Gaussian-like attributes to achieve their results, making it more difficult to explain and understand the underlying data structure.

Definition of PCA

Suppose that $\boldsymbol{x}$ is vector of n variables, $x_1, x_2, \ldots x_n$, of which the variance structure is of interest. PCA then calculates a new set of variables $\boldsymbol{Z} = Z_1, Z_2, \ldots Z_n$ which are linear combinations of the initial variables. That is, for some set of coefficients $a_{i,j}$:

$$Z_i = \sum_{j=1}^{n} a_{i,j} x_j$$

which can also be written as

$$\boldsymbol{Z} = \boldsymbol{A}^T \boldsymbol{x}$$

where the superscript T denotes matrix transverse. The following derivation of each of the new variables (or "principal components") is based upon a geometric approach and seeks the coefficients of the orthogonal transformation $\boldsymbol{A}$, such that the resultant principal components are uncorrelated [28, 76].

Consider the first variable

$$Z_1 = \sum_{j=1}^{n} a_{1,j} x_j$$

Choosing $\mathbf{a}_1 = (a_{11}, a_{12}, \ldots a_{1n})$ to maximise the variance of the principal component Z_1 then gives the variance as

$$\begin{aligned} var(Z_1) &= E[Z_1^2] - E[Z_1]^2 \\ &= E[\boldsymbol{a}_1^T \boldsymbol{x} \boldsymbol{x}^T \boldsymbol{a}_1] - E[\boldsymbol{a}_1^T \boldsymbol{x}] E[\boldsymbol{x}^T \boldsymbol{a}_1] \\ &= \boldsymbol{a}_1^T (E[\boldsymbol{x}\boldsymbol{x}^T] - E[\boldsymbol{x}] E[\boldsymbol{x}^T]) \boldsymbol{a}_1 \\ &= \boldsymbol{a}_1^T \boldsymbol{\Sigma} \boldsymbol{a}_1 \end{aligned}$$

where $E[\cdot]$ is the mathematical expectation of an expression and $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{x}$. Using the constraint that $\boldsymbol{a}_1$ is of unit length (i.e. $\boldsymbol{a}_1^T \boldsymbol{a}_1 = |\boldsymbol{a}_1|^2 = 1$) then gives that $var(Z_k) = \boldsymbol{\lambda}_k$ where $\boldsymbol{\lambda}_k$ is the kth largest eigenvector of $\boldsymbol{\Sigma}$. To satisfy this

constraint, a typical approach is to find the stationary value of:

$$f(\boldsymbol{a}_1) = \boldsymbol{a}_1^T \boldsymbol{\Sigma} \boldsymbol{a}_1 - \nu \boldsymbol{a}_1^T \boldsymbol{a}_1 \tag{1.18}$$

where ν is a Lagrange multiplier. Differentiating equation (1.18) with respect to $\boldsymbol{a}_1$ then gives:

$$\boldsymbol{\Sigma} \boldsymbol{a}_1 - \nu \boldsymbol{a}_1 = 0$$

which can be re-written using the $(n \times n)$ identity matrix $\boldsymbol{I}_n$ as

$$(\boldsymbol{\Sigma} - \boldsymbol{I}_n \nu)\, \boldsymbol{a}_1 = 0$$

For non-zero solutions, $\boldsymbol{a}_1$ is an eigenvector of $\boldsymbol{\Sigma}$ with ν as its eigenvalue. To determine the maximal variance of $\mathbf{a}_1 \boldsymbol{x}$, the largest eigenvector of $\boldsymbol{\Sigma}$, $\lambda_1 = var(\boldsymbol{a}_1 \boldsymbol{x})$, is then needed.

Generally, the kth principal components of of the initial variables $\boldsymbol{x}$ is $\mathbf{a}_k^T \boldsymbol{x}$ and $var(\boldsymbol{a}_k^T \boldsymbol{x}) = \lambda_k$ for λ_k the kth largest eigenvalue of $\boldsymbol{\Sigma}$ with corresponding eigenvector $\boldsymbol{a}_k$.

1.4.2 Linear discriminate analysis

Linear discriminate analysis (LDA) is a supervised multivariate technique as the classification of each sample is known prior to the analysis. By statistically modeling the properties of each of the experimental classes, it is possible to generate a vector which best separates the known groups, as well as predicting the classification of new samples using this decision boundary. LDA is similar to PCA in that it uses linear combinations of the sample vectors (i.e. the bin intensities of each spectrum) to create these rules, however, no transformation is applied to the data. In this thesis, Fisher's linear discriminant is used. Here, the model created is simply the linear combination of variables which minimises within-group variation and maximises between group variation [53]. This criterion is also written as maximising the ratio of between-class to within-class variances [76].

Model validation

An issue which occurs with supervised statistical methods is *over fitting*. This is when the model created by the analysis is overly complicated and extracts the sample noise as an important classification feature, rather than the true sample properties. Alternatively, the model created may be too simple and fail to extract classification features [76].

To test for over fitting, models are tested using test-data and cross-validation. This thesis uses the simple method of leave-one-out (LOO) cross-validation. To analyse models using LOO classification, n data subsets are created by removing a single sample from the dataset (for each sample in turn), then the analysis performed upon these subsets. If any of the models created by the data subsets are radically different from the full data, then the model is over fit and should be discarded.

Definition of LDA

This definition assumes that the samples are split into two experimental groupings, as used in this thesis. LDA may, however, be extended into multiple class problems, as described in many statistical texts [53, 76]

Consider a group of n samples split into two groups of n_1 and n_2 samples ($n_1 + n_2 = n$). To describe the separation between the two classes, it is necessary to find the vector $\boldsymbol{w}$ which maximises the ratio

$$J_F = \frac{|\boldsymbol{w}^T(\boldsymbol{m}_1 - \boldsymbol{m}_2)|^2}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}} \tag{1.19}$$

where $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ are the group means and $\boldsymbol{S}_w$ is the within-class covariance, given as

$$\boldsymbol{S}_w = \frac{1}{n-2}\left(n_1\boldsymbol{\Sigma}_1 + n_2\boldsymbol{\Sigma}_2\right) \tag{1.20}$$

where $\boldsymbol{\Sigma}_i$ are the maximum likelihood estimates of the covariance matrices of each class. Turning points (i.e. maxima) of equation (1.19) can be found by differentiating and

equating to zero. That is, solving the equation

$$\frac{\boldsymbol{w}^T(\boldsymbol{m}_1 - \boldsymbol{m}_2)}{\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}}\left[2(\boldsymbol{m}_1 - \boldsymbol{m}_2) + \frac{\boldsymbol{w}^T(\boldsymbol{m}_1 - \boldsymbol{m}_2)}{\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}}\boldsymbol{S}_w\boldsymbol{w}\right] = 0 \tag{1.21}$$

Since the LDA classifier is the line that best separates the classes, only the direction of the vector $\boldsymbol{w}$ is needed. Thus, equation (1.21) can be reduced to

$$\boldsymbol{w} \propto \boldsymbol{S}_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2) \tag{1.22}$$

as $\boldsymbol{w}^T(\boldsymbol{m}_1 - \boldsymbol{m}_2)/\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}^T$ is a scalar value.

To derive an allocation rule to classify a new sample, $\boldsymbol{x}$, it is necessary to specify a threshold, w_0 such that

$$\boldsymbol{w}^T\boldsymbol{x} + w_0 > 0 \tag{1.23}$$

Using the nearest class-mean (i.e. class centroid) classifier for the two class problem (see Webb [76]), equation (1.23) can be written as:

$$= \boldsymbol{S}_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2) \tag{1.24}$$

and

$$w_0 = -\frac{1}{2}(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T\boldsymbol{S}_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2) - \log\left(\frac{p(\omega_2)}{p(\omega_1)}\right) \tag{1.25}$$

where $p(\omega_i)$ is the (known) prior probability of membership in class i.

Chapter 2

NMR Spectral Analysis

This chapter describes the whole spectral analysis of NMR using relative standard deviation (RSD). Firstly, RSD is introduced as a suitable tool to evaluate the reproducibility and robustness of spectral data. The principles and applications are discussed, along with some limitations. Spectral denoising and noise profiles of NMR spectra are also introduced and discussed as part of evaluating experimental structure of both line and surface spectra. Finally, examples of spectral RSD are presented to illustrate the many potential uses in the field of metabolomics.

2.1 Spectral data quality

When conducting a metabolomics study, the acquisition of data of sufficiently high quality is a priority. To correctly focus the analysis upon the biological phenomena being studied, it is essential to understand and, potentially, minimise the spectrum-wide variation arising from technical sources as well as inter-individual metabolic variation within each class. Without this, the interpretation of results can be ambiguous, misleading and in the worst case, false [33, 59].

2.1.1 Relative standard deviation

Spectral variation is difficult to measure, however, since NMR spectra have two distinct types of measurement error, one additive and one multiplicative [56] (described in section 1.3.7 and shown in figure 2.1). This then has the effect that the standard deviation of any data point is related to its intensity, limiting the usefulness of any investigation into variation of spectral intensity. Relative standard deviation, RSD (also termed coefficient of variation, CV), is a different approach for characterising measurement variability. The RSD of a point is simply defined as the standard deviation of the data at that point divided by the corresponding mean value, multiplied by 100 (equation (2.1)). RSD is a dimensionless quantity and allows the easy comparison of the variability of data with vastly different mean values, but also limits the effects of the multiplicative error. Notice that RSD is undefined at $\mu = 0$.

$$\text{RSD(x)} = \frac{\sigma(x)}{\mu(x)} \times 100 \tag{2.1}$$

The application of RSD in metabolomics is not new, for example, Keun *et al.* [33] reported RSD values for the measurement of citrate, taurine and hippurate in urine samples acquired by NMR. However, the RSD of an entire spectra set would also be advantageous as this would describe the reproducibility of the whole metabolomics dataset. To date, there has been limited use of RSD for capturing "spectrum-wide" variability in metabolomics studies, since the results were distorted by the inclusion of spectral noise in the analysis [8, 59, 63]. The inclusion of spectral noise (i.e. baseline with no signal) misrepresents the results, since by its very nature noise provides a unique set of values for each spectrum. It is further compounded by the fact that noise in NMR and mass spectrometry (MS) spectra fluctuates around zero, since a limitation of RSD is that it is extremely sensitive to small fluctuations when the mean value is near zero. However, spectrum-wide RSD has great potential to provide practical benchmarks in all fields of metabolomics. For example, spectral RSD of an established method can provide a bench-

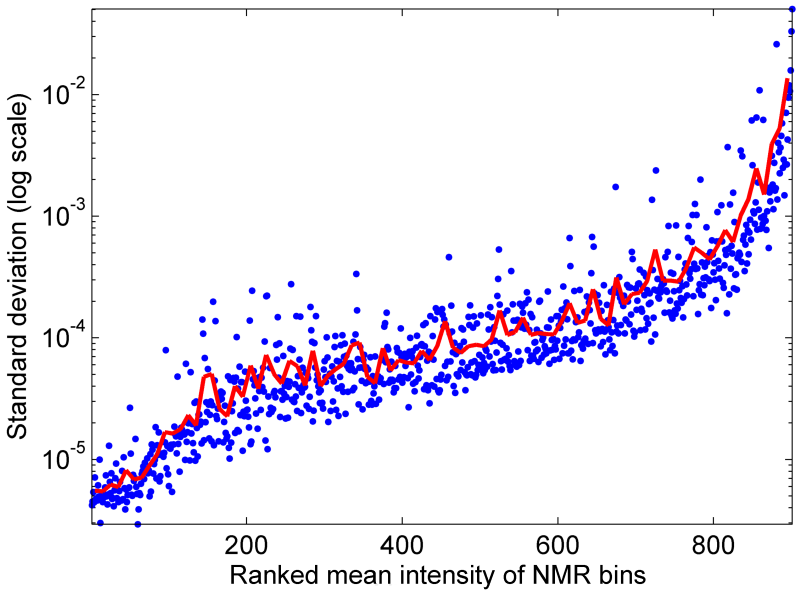


Figure 2.1: Plot of (log) standard deviation versus the ranked mean bin intensity of 27 mussel tissue samples. Here it can be seen that the standard deviation increases as the mean bin intensity increases. The solid red line represents a trendline with running median of 10 bins.

mark for developing new methodologies or for applying the original method to different biological sample types. In addition, established spectral RSD values can provide new researchers in metabolomics with "target values" for achieving high quality studies.

2.1.2 Noise estimation in NMR spectra

Since the inclusion of noise in a spectrum wide analysis such as using RSD can produce artifacts or obscure true signal, it is advantageous to filter the spectra and remove the unwanted noise before any analysis. However, to de-noise an NMR spectrum successfully, it is necessary to be able to understand and approximate the unwanted signal for each experiment type.

Noise produced by an NMR spectrometer is not strictly uncorrelated Gaussian noise, since the effects of digitising the signal after sampling subtly alters the initially Gaussian distributed signal produced in each of the real and imaginary channels of the spectrometer[18]. However, as shown in figure 2.2, a model Gaussian distribution (solid red line) is an appro-

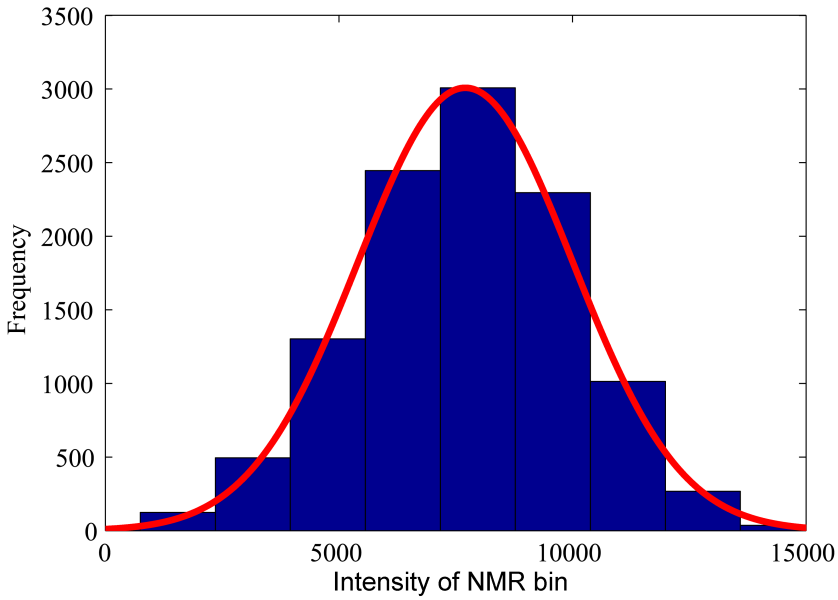


Figure 2.2: Histogram of noise present in a Fourier transformed 1D NMR spectrum. The solid red line represents the fit of the data to a Gaussian distribution.

priate approximation of the distribution of the noise. This justifies the assumption that for 1D data, the majority of the noise is encompassed by 3 times the standard deviation away from the mean value of the baseline (i.e. about zero) as is commonly assumed by many NMR papers[16, 40].

For J-resolved (JRES, see section 1.3.3, p.18) data, a 2D NMR experiment, the two channels of signal acquisition are combined in a non-linear manner to create the magnitude mode spectrum. Since each of the channels contains the Gaussian-like noise of the 1D spectra, it is then likely that the JRES spectra exhibit Rayleigh-like noise (see appendix C). A Rayleigh distribution is right skewed and non-negative and it is often used to described data that is a result of calculating the magnitude of two uncorrelated Gaussian distributed variables of equal variance and zero mean. A histogram of the noise found in a JRES spectrum is shown in figure 2.3, where both Gaussian (solid red) and Rayleigh (dashed green) distributions have been fit to the data. It can be seen that both distributions model the spread of the data well, but differ slightly in the estimation of the tails, as the Gaussian estimation under estimates the right tail. Thus using 3 times

the standard deviation of the data provides a useful noise measure, but is susceptible to positive 'spikes' in the noise.

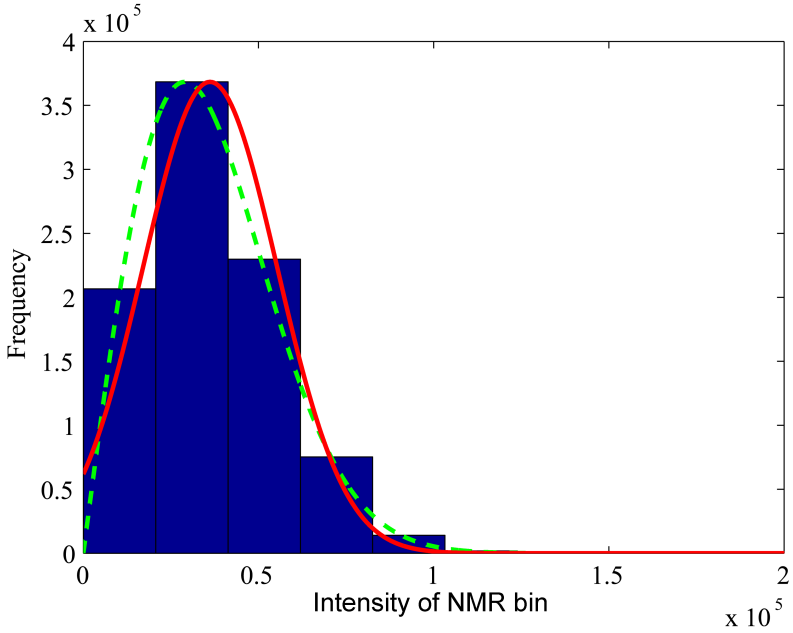


Figure 2.3: Histogram of noise present in a Fourier transformed JRES NMR spectrum. The solid red line represents the fit of the data to a Gaussian distribution and the dashed green line represents the Rayleigh fit. Both distributions give a fair representation of the data, however, the Rayleigh distribution has a better fit of the tails of the data.

A histogram of the noise found in a projected J-resolved (pJRES, see section 1.3.3, p.18) spectrum is shown in figure 2.4. Here the noise profile strongly resembles the profile of the 1D rather than the 2D JRES spectra. Both Rayleigh (green dashed) and Gaussian (red solid) distributions have been fit to the noise, where clearly, the Gaussian distribution best fits the data. Again, the standard of 3 times the standard deviation of the data is a good measure for estimating the noise of this experiment type.

Whilst using a Gaussian approximation for each of the experiment types may give a fair approximation of the noise; simply using 3 times the standard deviation leaves noise spikes in the spectra unaccounted for, especially for the intact JRES data. This may affect the accuracy of any further analysis of the data by misclassifying noise as signal, for example. Hence, a method to further discriminate between noise and signal is needed. A simple

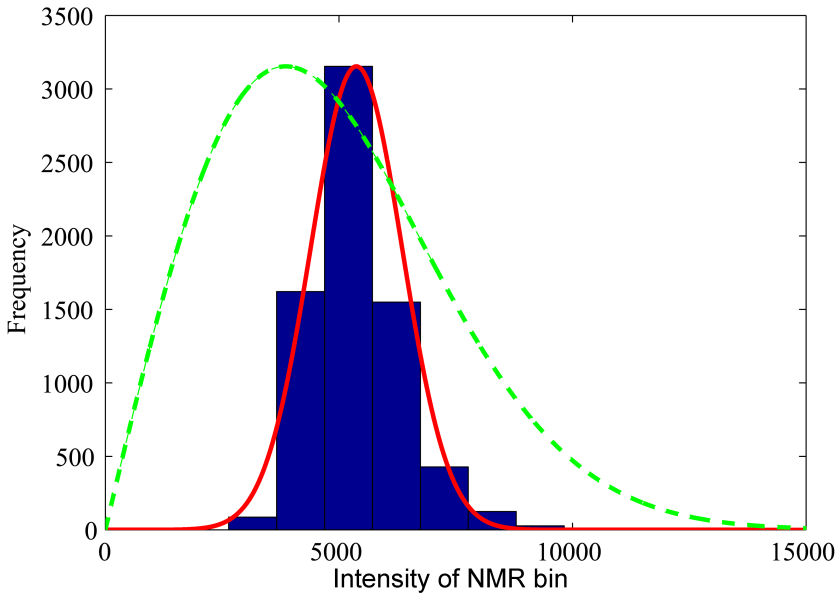


Figure 2.4: Histogram of noise present in a Fourier transformed pJRES NMR spectrum. The solid red line represents the fit of the data to a Gaussian distribution and the dashed green line represents the Rayleigh fit. Clearly, the Gaussian distribution provides the more accurate fit.

way of further filtering the data to remove the effects of noise spikes is to classify a bin as containing signal if and only if its intensity is above the established "noise threshold" for every spectrum in the given data set. This also has the further advantage of limiting the effects of peak shifting.

The denoising algorithm

To calculate a noise threshold, for the 1D and pJRES data, each binned spectrum was divided into 32 sections and the standard deviation of each section calculated. The noise level for a given spectrum was estimated as 3 times the smallest standard deviation of these 32 sections, as reported previously [16, 40]. For 2D JRES data, a noise surface was calculated using a similar approach in that a noise level was estimated (exactly as above) for each of the 128 spin-spin coupling increments. Next the 2D JRES spectra and corresponding noise surfaces were concatenated into 1D row vectors. Bins containing only noise were then removed from each of the spectra. This is the standard denoising

algorithm used throughout this thesis.

2.1.3 Spectral RSD

After de-noising the spectra, it is then a simple matter to calculate the RSD of each remaining "signal bin" in the data. This procedure generates a distribution of RSD of several hundred values, depending on spectral quality, sample type and experiment. Interpretation and analysis of the data is therefore dependent on the statistical analysis, such as median and range values (see figure 2.5 for an example distribution). Appendix B contains some example statistics for data sets discussed in this thesis.

Presentation of the data is also extremely important, as the RSD of each data set must be easily compared with each other, as well as summarising the values. As shown in figure 2.5, there is a variety of methods available emphasising different aspects of the distribution. Here, 27 ^{1}H 1D NMR spectra acquired from muscle samples from the marine mussel *Mytilus galloprovincialis* have been de-noised and their spectral RSD calculated. Figure 2.5A illustrates each of the 901 signal bins ranked by their mean intensity versus their RSD. Clearly, the dependence upon the intensity of the bin has been removed. Figure 2.5B shows the histogram of the RSD data where the large right tail and skewed nature of the distribution can be seen. Part C shows the data redrawn as a boxplot. The boxplot is the method used within this thesis to portray the RSD distributions, since this method facilitates visual comparison of RSD distributions from multiple datasets, showing lines at the lower quartile, median and upper quartile values, whiskers to display the range of the remaining data, and outliers as individual data points. Outliers correspond to RSD values that are 1.5 times the interquartile range (or more) below the lower quartile, or 1.5 times the interquartile range (or more) above the upper quartile. Throughout this thesis the boxplots show RSD from 0-100% to enable comparisons across all datasets.

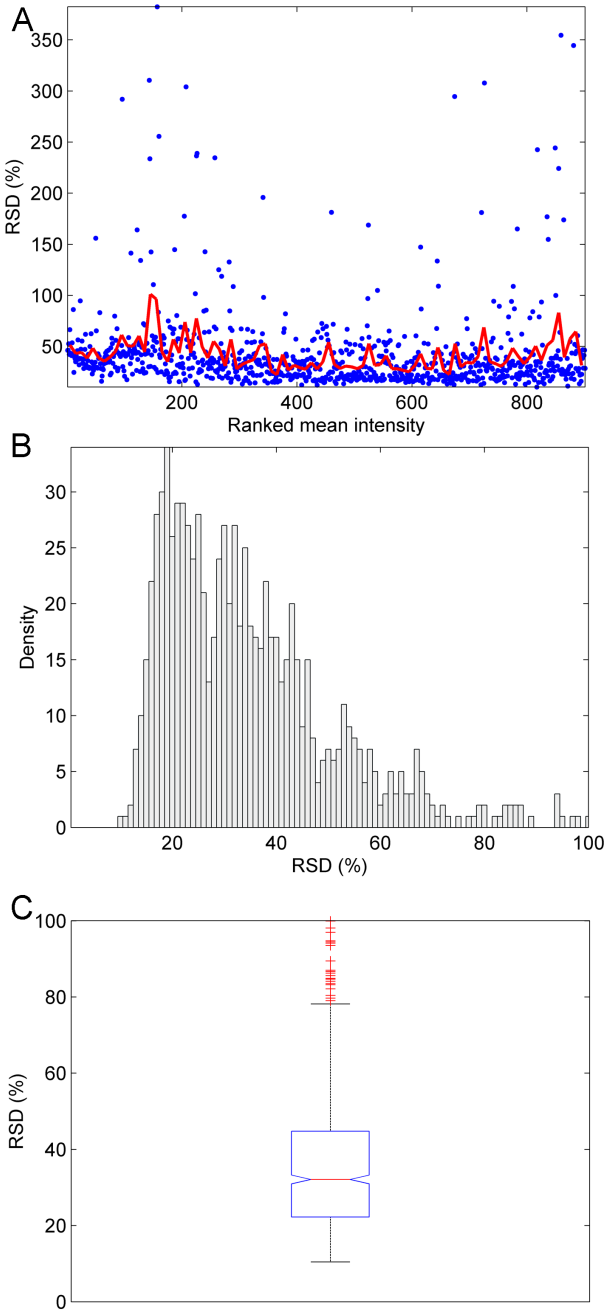


Figure 2.5: Presentation of spectral RSD values. (A) Following removal of bins that contain only noise, RSDs are calculated for each remaining bin and then ranked according to bin signal intensity. A trendline (with running median of 10 bins) highlights that the RSDs are largely invariant to signal intensity. (B) Histogram of RSD values showing a right-skewed distribution. (C) Boxplot of RSD values which summarises succinctly the lower quartile, median and upper quartile values, whiskers to display the range of data, and outliers as individual data points. Both the boxplot and histogram have been clipped to the region 0-100% RSD.

Limitations

As with any statistical technique, RSD suffers from constraints and limitations of usage. For example; it is known that the confidence interval of the RSD is proportional to the inverse of the number of samples present[30]. Hence during experimental design, enough samples must be acquired to reduce the confidence interval to appropriate levels. For example, to achieve an 80% incidence of a confidence interval of 5%, 11 samples are necessary[30].

An issue of concern also arises from the fact that each RSD distribution can also be considered a point in multi-dimensional space, with the RSD of each bin corresponding to a single dimension. Analysing data in multivariate space can be very demanding, since to ensure that the root mean square (RMS) error of the density estimation is lower than a given threshold, the number of samples required increases greatly as the number of dimensions increases. For example, to ensure that RMS error at zero is less than 0.1 when estimating a standard multivariate normal density it is recommended that at least 4 samples are required for 1 dimension; 768 samples for a 5 dimensional space and around 842,000 samples for 10 dimensional data [60]. Clearly, for several hundred variables the number of samples would be even greater and this level of precision can never be realistically reached with NMR metabolomics.

Experimental ramifications

Both of these issues illustrate the need for large numbers of samples in any given metabolomics investigation, to limit the errors. Unfortunately, this is not always feasible, with limitations including such factors as: the availability and quantity of samples; experimental facilities; time; man-power and budget constraints. However, if the experimenter is aware of these issues, the limitations of the techniques can be allowed for during experimental design.

2.2 Examples of spectral RSD

This section provides examples of the use of spectral wide RSD as a measure of variance across multiple metabolomics datasets. The focus is on denoised binned spectra from ^{1}H 1D NMR spectroscopy as this experiment encompasses some of the currently most widely used approaches in metabolomics; but other NMR and MS data is also included, including both pJRES and intact JRES spectra. Here the RSD is displayed succinctly in the form of boxplots and the applications and usage of the results discussed. This chapter is based upon work published in *The Analyst* [46].

2.2.1 Data acquisition and comparison

NMR spectra were acquired from multiple datasets [36, 72, 65, 22, 61, 71, 73, 47, 5, 9] and processed using standard methods, including Fourier transformation, phasing, baseline correction and calibration with either XWINNMR (Bruker), TopSpin (Bruker) or ACD/1D NMR Processor (Advanced Chemistry Development) software. All 1D ^{1}H NMR spectra as well as JRES spectra and their 1D skyline projections were processed using ProMetab[71] in Matlab using standard methods (see relavent citations for specific details). For the MS data, Fourier transform ion cyclotron resonance (FT-ICR) mass spectra of fish liver extracts were initially processed as described in Southam et al.[61], yielding a list of peaks between 70-500 Da for each sample. For each list, the total peak intensity was normalised to unity and the peaks were sectioned into 430 bins of width 1 Da.

After processing and de-noising, the spectral RSD of each data set was calculated. As established in section 2.1, the RSD removes the relationship between bin intensity and stability that results in only examining the standard deviation of the data. As shown in figure 2.5A, it is now possible to see that most bins have a similar RSD value, with only a small percentage of points exhibiting high values. Since the distribution of the RSD is clearly non-Gaussian due to the high skewness, the analysis and comparison of multiple RSD data sets becomes difficult. This is because most statistical methods rely on the

Gaussian nature of the data to achieve their results. Hence to analyse the different data sets, non-parametric tests are necessary to establish whether the RSD distributions are different. In this chapter, the statistical tools used to analyse the data are the Wilcoxon rank sum test for comparing two distributions and the Kruskal-Wallis test for multiple distributions. All RSD statistics for each of the data sets can be found in tables B.2 and B.3.

2.2.2 Applications of spectral RSD

Technical replicates

An obvious use for the RSD of technical (or analytical) replicates is to evaluate spectral quality in terms of the experimental methods. Here, RSD can be used to compare spectra obtained using different experimental methods such as sampling protocols, metabolite extraction and analysis procedures or data processing strategies. It is particularly useful when comparing a new method to an existing one. Several example applications are discussed below.

Optimising metabolite extractions

Two protocols were compared for extracting metabolites from a chub liver (*Leuciscus cephalus*), 6% perchloric acid (n=10) and 2:1 methanol:water (M:W; n=5). The two RSD distributions derived from 1D NMR spectra are significantly different ($p < 0.001$; figure 2.6), and highlight the considerably higher reproducibility of the M:W extraction since the median RSD of the M:W method is only 4.6%, whilst the perchloric acid extraction exhibits a median RSD of 20.6%. This is consistent with observations from the original study that NMR spectra of perchloric acid extracts showed considerable pH-induced peak shifting [36].

Sampling, extraction and matrix effects

Different sample types, such as biofluids, cells or tissues, are typically collected and the metabolites extracted using protocols of differing complexities. Also, the resulting extracts will be present in different sample matrices. Together these factors will affect the reproducibility of the metabolomics data. NMR spectra of four disparate sample types were compared: urine from a dog that was collected by free-catch, divided into five samples and prepared by simple buffer addition[72]; a flask of acute myeloid leukaemia cells (cell line K562) that was divided into five samples, centrifuged, washed and extracted using methanol:chloroform:water (M:C:W)[65]; adductor muscle from a Mediterranean mussel (*Mytilus galloprovincialis*) that was rapidly dissected, frozen, homogenised, divided into six and then extracted using M:C:W[22]; and liver from a 3-spined stickleback, (*Gasterosteus aculeatus*) that was processed in an identical manner to the mussel tissue, producing six samples. Figure 2.6 shows the differences between the four RSD ($p < 0.001$) and confirms that the simplest collection and extraction procedure used in conjunction with the simplest aqueous matrix results in the most reproducible data (median RSD of 1.6% for urine). The rapid dissection, freeze clamping, homogenisation and extraction of more complex tissue samples increases the median RSD to 3.4% (stickleback) and 6.1% (mussel). The preparation of cell extract replicates, requiring several minutes of centrifugation and washing prior to quenching the metabolites, yielded the least reproducible data (median RSD of 14.0%). These values serve as valuable benchmarks.

Comparison of analytical approaches

Three datasets were recorded using different NMR experiments, comprising traditional 1D 1H, 1D projections of JRES spectra, and intact 2D JRES spectra. In all cases, the same five replicates of a European flounder (*Platichthys flesus*) liver extract were analysed. The boxplots corresponding to the three sets of RSD values illustrate that 1D NMR spectroscopy yields the most reproducible data (median RSD of 3.1%), the intact 2D JRES spectra show the highest variability (19.8%), and the pJRES data is intermediate between

the other two with a median RSD of 12.5% ($p < 0.001$; figure 2.6). This can be explained in part by considering the number of scans used for each type of experiment: 100 averaged transients for each more quantitative 1D spectrum but only 16 for each JRES spectrum, yielding higher accuracy in the 1D measurement. In a separate study, technical variability of metabolite levels in flatfish liver extracts were compared using 1D NMR spectroscopy (one flounder liver, median RSD of 3.1%) and direct infusion FT-ICR mass spectrometry (one dab liver, median RSD of 13.1%). Although these are not identical samples, they are the same matrix from closely related species that were collected, quenched and extracted using identical protocols. The superior spectral reproducibility for the 1D NMR analysis ($p < 0.001$; figure 2.6) is not surprising given the excellent analytical precision typically observed when using NMR.

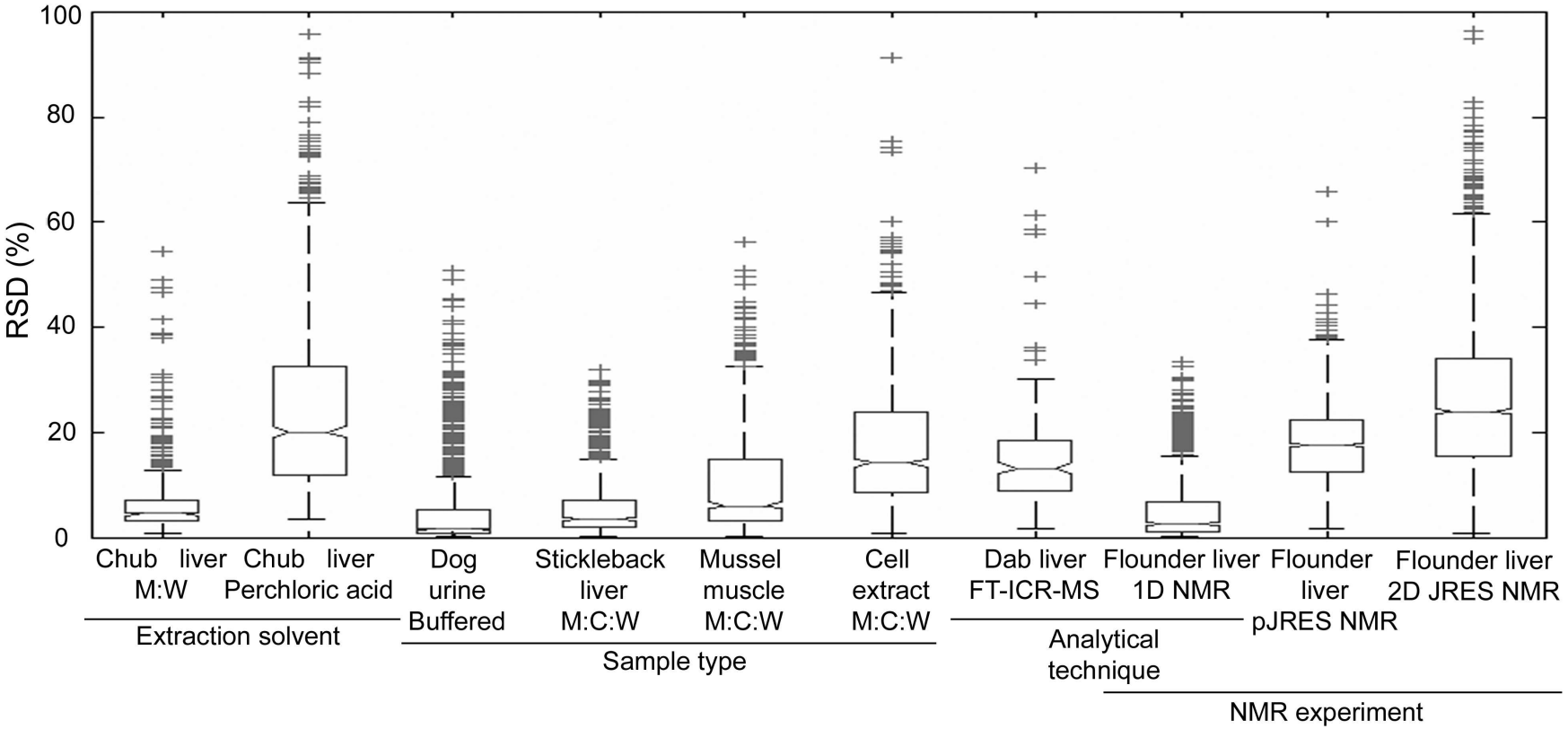


Figure 2.6: Boxplots of RSD derived from technical replicate spectra for 10 independent datasets, as described in the main text. These datasets facilitate comparisons of the reproducibilities associated with two solvent extraction methods, four types of biological sample, two analytical techniques, and three types of NMR experiment. Key: M:W = methanol:water extraction; M:C:W = methanol:chloroform:water extraction

2.2.3 Applications for RSD of inter-class variation

Variation between individuals in a single class or group includes genuine biological inter-individual differences as well as technical variation. Quantifying this inter-individual variation can be useful for optimising experimental design, as illustrated below.

Comparison of samples from a single species

It is commonly believed (but rarely shown) that inter-individual metabolic variability is biofluid or tissue dependent. This is due to tissues being under greater homeostatic control than biofluids, so it is predicted that metabolic variability increases from tissue to plasma to urine - the latter being particularly susceptible to diet and lifestyle. Figure 2.7A compares the RSDs derived from NMR datasets of three sample types, all derived from 8 separate marine invertebrates (red abalone, *Haliotis rufescens*)[71]. The RSD distributions for foot muscle, digestive gland and haemolymph (blood) are significantly different ($p < 0.001$) with haemolymph showing the greatest inter-individual variability, consistent with the prediction of homeostatic control. The RSDs of NMR spectra of rat brain and plasma, obtained from five individuals[73], again show that the biofluid exhibits somewhat greater variability ($p < 0.001$; figure 2.7B). Interestingly, these inter-individual median RSDs of 7.2% (brain) and 8.0% (plasma) are considerably smaller than for fish and marine invertebrate tissues and biofluids (that range from 16.0-58.4%), it is likely that this reflects the conserved metabolism in laboratory-raised Sprague-Dawley rats, which were all male adults, of mass 325-375 g and were individually housed in environmentally-controlled chambers under a 12:12 hr light:dark cycle. As a further example, figure 2.7C shows the variance within NMR spectra of testis, urine and plasma from 7 fathead minnow (*Pimephales promelas*). The tissue again exhibits the lowest RSDs (median of 29.4%), while the two biofluids have similar inter-individual metabolic variability ($p = 0.263$). It should be noted that while tissues are generally under greater homeostatic control, some exhibit considerable heterogeneity (e.g. tumour tissue) and therefore will produce larger than expected biological variation. Overall these results can help to guide experimental

design, for example by identifying which sample type yields the lowest inter-individual variability.

Comparison of different species

The same sample type can also be compared across different species, which may be particularly useful for determining the feasibility of studying non-model organisms; i.e. to confirm that the inter-individual metabolic variability for that species is sufficiently low to produce interpretable data. Figure 2.7D compares the RSDs from NMR spectra of rat urine (n=5) and fathead minnow urine (n=7). All samples were buffered and processed using identical procedures. However, the RSDs are quite different ($p < 0.001$), potentially reflecting differing biochemistries of these organisms.

Optimisation of organism "husbandry"

Inter-individual variation can be used to assess protocols for the culturing or husbandry of organisms, as the condition in which animals are kept will have a large impact on their metabolic variability. This is particularly true for metabolomics studies of wildlife when organisms are housed in "foreign" laboratory environments. Inter-individual variation was compared for mussels (*M. galloprovincialis*) collected directly from the field (n=14), versus animals collected from the field and then maintained in a controlled laboratory environment for 48 hr (n=14)[22]. The median RSDs derived from NMR spectra show that the laboratory introduces a small but significant increase in the metabolic variability in two tissues, the adductor muscle ($p < 0.05$, Wilcoxon) and mantle ($p < 0.001$) as shown in figure 2.7E. It was concluded that direct sampling from the field was preferable, as discussed in the original paper[22].

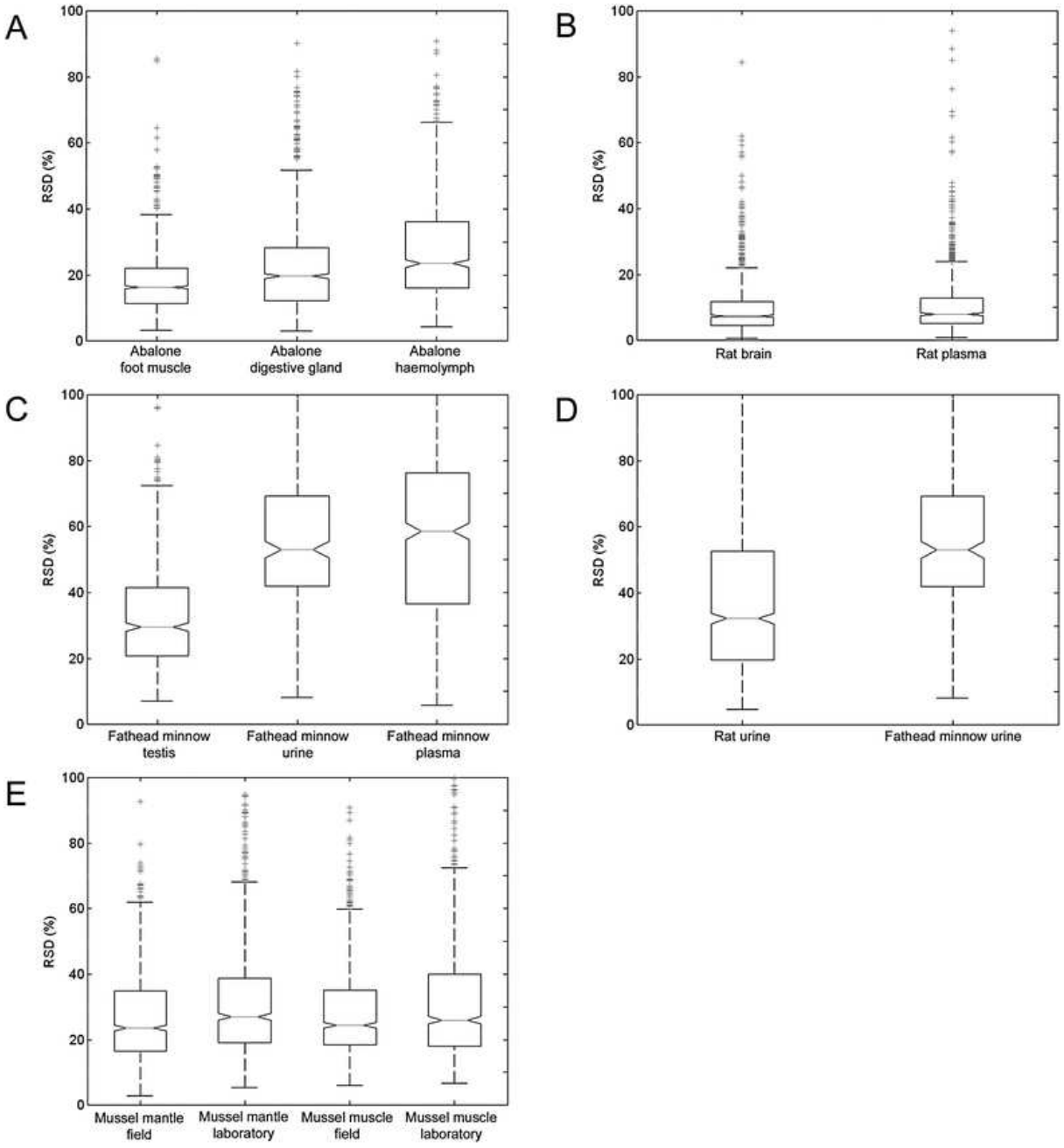


Figure 2.7: Boxplots of RSD derived from several NMR datasets that facilitate comparisons of inter-individual metabolic variation across classes. (A) Three types of biological sample from abalone shellfish. (B) Tissue and biofluid samples from rat. (C) Three types of biological sample from fathead minnow. (D) Urine samples from two different species. (E) Comparison of husbandry and sampling techniques for two tissues from marine mussels.

2.2.4 Applications for RSD of inter-class variation

The goal of most metabolomics experiments is to discriminate between sample classes based upon their metabolic compositions. It is therefore important and useful to confirm that technical variation is small relative to the inter-individual variation within each class, prior to interpreting the biological variation. This is clearly illustrated in the following three examples.

Effect of environment on fish liver metabolome

Flounder were collected from the Rivers Alde (n = 19; clean) and Tyne (n = 18; more polluted) to determine the effects of different environments (and potentially different genetic make-ups) on the liver metabolome. 14 RSDs were derived from NMR spectra of individuals from each site as well as from technical replicates of one individual (figure 2.8A). Clearly, technical variation is considerably lower than the biological variation between fish from either site. Furthermore, the significant difference ($p < 0.001$) between the inter-individual variations for the two sites suggests that the two fish populations are responding differently to their environments, with the control fish from the clean site showing greater metabolic variability.

Effect of drug treatment on cellular metabolome

K562 leukaemia cells were treated with medroxy progesterone acetate. Figure 2.8B illustrates the RSDs derived from NMR spectra of both treated (n=12) and untreated (n=12) flasks of cells, as well as from five technical replicates. Again the technical replicates show the least metabolic variation. For this study, although there is a significant difference between the inter-individual variations ($p < 0.01$), the cells treated with the drug exhibit greater metabolic variability.

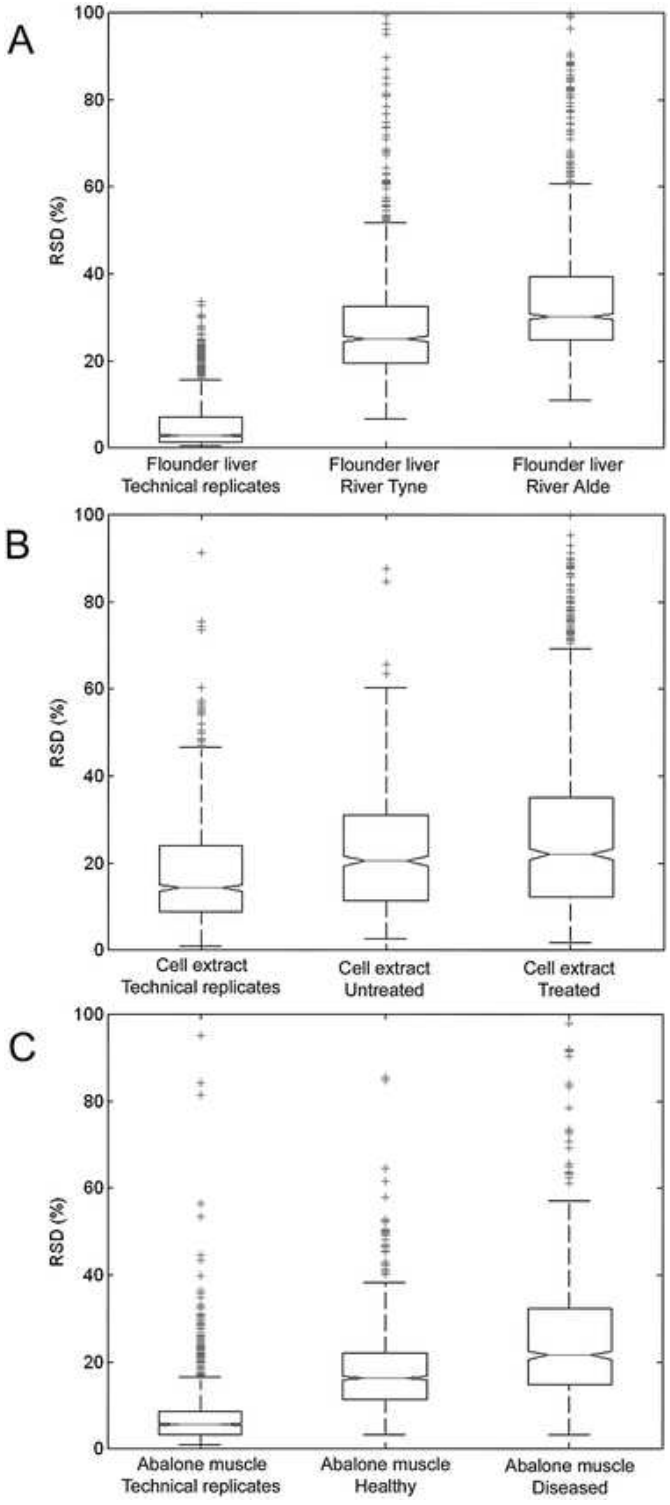


Figure 2.8: Boxplots of RSD derived from several NMR datasets that compare technical variation (in one sample) to inter-individual metabolic variation across classes. (A) Effect of sampling site on flounder liver extracts. (B) Effect of drug treatment on K562 leukaemia cell lines. (C) Effect of withering syndrome disease on abalone adductor muscle. In all cases, technical variability is shown to be smaller than inter-individual variability.

Effect of disease on metabolome

NMR spectra of foot muscle were obtained from healthy abalone shellfish (n=8), animals with withering syndrome (n=5), as well as technical replicates from one individual (n=5)[71]. Figure 2.8C shows that the technical replicates have the smallest variation (median of 5.4%), confirming that variability arising from the experimental protocol is minor compared to the biological differences. Furthermore, the diseased class exhibits a significantly higher inter-individual variation than the healthy controls ($p < 0.01$). As withering syndrome is a chronic disease that progressively degrades the metabolic condition of the shellfish over many months, this high variation can be rationalised in terms of the animals being in various stages of the disease. Overall, considering just these few examples, it suggests that no "golden rule" exists for the control class exhibiting higher variability.

2.3 Conclusions

RSD is an underused tool for assessing the reproducibility of metabolomics datasets, which has the potential for many useful applications. In this chapter, ten NMR and MS datasets have been analysed in terms of technical variation or inter-individual variation, spanning a variety of sample types including fish, invertebrates, mammals and a model cell line. RSD has been shown to a useful and versatile tool for comparing data that would otherwise be exceedingly difficult to compare and contrast as well as assessing data quality and improving experimental design. However, RSD suffers from the limitation of wide confidence intervals with small sample sizes, limiting accuracy. It is then concluded that this benchmark could be of considerable value to both existing and new practitioners in the field and for both the interpretation of technical and biological variation. It is therefore recommend that a database of RSD values be established for a wide variety of metabolomics datasets for future use.

RSD is also used for spectral analysis in chapter 3 of this thesis to help illustrate the effects of variance scaling transformations and then again in chapter 4 to compare the processing methods used in the acquisition of JRES NMR spectra.

Chapter 3

Variance Scaling

The variance structure of NMR spectra is an often overlooked, yet vitally important part of the data. This is since many commonly used data mining tools such as principal component analysis (PCA); partial least squares discriminant analysis (PLS-DA) and many other tools which rely extensively on variance are used to analyse metabolomic data sets. It is therefore important to assess that the variance structure of the data is appropriate for use with the analysis tool. Unfortunately, NMR spectra violate the basic assumption of homoscedasticity (constant variance) that many mathematical tools assume. This is because the variance of spectral bins between multiple experiments with large intensities is larger than the variance of bins with small intensities (see section 1.3.7). However, many data processing methods exist to help transform NMR; and other metabolomics data sets; into formats more suitable for multivariate analysis.

In this chapter, the use and application of transforms and scaling methods for improving classification of samples after multivariate analysis are discussed. Firstly, several common methods are defined: autoscaling, Pareto scaling and the generalised logarithm (glog) transformation. Parameter choices and a modification for the glog transformation are also discussed. A scaling intercomparison is also presented, detailing the effectiveness of the three methods (along with the unscaled data) upon both 1D and JRES NMR spectra of three disparate data sets.

3.1 Scaling methods

Common processing methods in metabolomics include mean centering, autoscaling, Pareto scaling, range scaling, VAST scaling[31], log transformations, and power transformation [69]. The focus of this thesis is on three commonly used techniques applied to NMR spectra: 'autoscaling'; 'Pareto scaling' and the 'generalised logarithm'.

3.1.1 Autoscaling

Autoscaling is a processing technique in which the variance of the intensity of each bin in the NMR spectrum is scaled to unity. For comparison between the other methods, the mean of each variable is then set to zero. This is a severe transformation that removes any variance bias in the data, making the data more consistent with the assumptions of many multivariate tools. However, it also removes any relevant biological variance patterns relating to the experiment.

3.1.2 Pareto scaling

Pareto scaling alters each variable's intensity by the square root of the standard deviation of that variable, producing a data set where the variance changes from variable to variable, but the range of variance across each spectrum is much reduced from the initial, unscaled data. It is a much less severe scaling method than autoscaling and reduces the changes in the variance without fully removing it. As an unsupervised method, it does not allow for the fact that there are different types of variation which differ in their importance to the experiment. Unlike autoscaling, Pareto scaling also changes the dimensions of the experiment; since however, NMR spectra are dimensionless quantities, this has no bearing upon the datasets presented in this thesis.

3.1.3 The generalised logarithm

The generalised logarithm (glog) has also been investigated, but is not widely used [54, 55]. For each variable in the spectrum, the glog transforms the intensity at that point to a value dependent on both the original intensity and the value of a transform parameter. The equation for the glog transform is shown below in equation 3.1, where y represents the untransformed data, λ is the transform parameter, and z is the transformed data [51].

$$z = \ln\left(y + \sqrt{y^2 + \lambda}\right) \tag{3.1}$$

The glog is a transformation that was originally applied to microarray data and is based on the two-component error model [54]. Unlike autoscaling and Pareto scaling, the glog transform is a *supervised* transformation, initially requiring a parameter to be calibrated from a series of NMR spectra of 'technical replicates' [47]. These replicates must be recorded from one (often pooled) biological sample which is divided into multiple components, each of which is subject to independent sample preparation and NMR analysis. Typically, 4 or 5 technical replicates should be used to calibrate the parameters. Thus the glog transformation reduces the amount of variance within the data set that arises solely from technical sources. Hence, when the glog is then applied to a biological data set, it effectively reduces the amount of technical variance present, leaving the biological variance to dominate in any subsequent multivariate analysis.

3.1.4 The extended glog transformation

The extended glog (ex-glog) transformation, defined by equation 3.2, is a modification to the glog transformation[1]. The extension of a second parameter, y_0, provides primitive denoising by shifting the focus of the transformation away from the small, unwanted noise

[1]The conception of the extended glog transformation belongs to C. Ludwig and U. Günter. The refinement and application however, remains the author's work.

peaks and focusing upon the signal peaks produced by the sample.

$$z = \ln\left((y - y_0) + \sqrt{(y - y_0)^2 + \lambda}\right) \tag{3.2}$$

3.2 Parameter estimation of the glog transform

3.2.1 Estimation of λ

The λ parameter in the glog transformation is specific to each data set and hence must be found prior to using the transformation. To compensate for only the unwanted technical variance in the samples, the calibration must be performed on a set of technical replicates generated from a single pooled biological sample. The replicate spectra are processed in exactly the same manner as the biological data set, i.e. normalisation, compression regions etc (see section 1.3.6), to ensure all technical variance is accounted for when calibrating the glog parameters.

The calibration process used in this thesis has been published in Parsons et al.[47] and is based upon a maximum likelihood method proposed by Rocke and Durbin [55]. In this thesis, to avoid scaling artefacts arising from the change in scale between the untransformed and transformed variables, the Jacobian of the glog function is used as a scaling factor (as detailed in Rocke and Durbin [55]). In this analysis, however, an alternative scaling function is used to the one detailed by Rocke and Durbin. This maintains most of the properties of the Jacobian but is computationally more robust. It is shown in equation (3.3), where z_i represents the intensity of bin i and n is the total number of bins contained within the spectrum being scaled.

$$J = \exp\left(\frac{\sum_{i=1}^{n} \ln\sqrt{z_i^2 + \lambda}}{n}\right) \tag{3.3}$$

The parameter λ was optimised by minimising the variance, S, (equation 3.4) over k

technical replicates and all n bins in the Jacobian-scaled data vectors $w_j = z_j J$, giving a measure of all variance contained within the technical replicates.

$$S(\lambda) = \sum_{j=1}^{k} \sum_{i=1}^{n} (w_{ij} - \hat{w}_i)^2 \tag{3.4}$$

$\hat{w}$ is calculated as the mean spectrum of all scaled and transformed technical replicates, w_j . Minimising the variance S thus gives an optimal value for λ. The optimisation is achieved via the Nelder-Mead unconstrained non-linear minimization routine in the MATLAB optimisation toolbox. The optimised λ value was then used to transform the binned intensities of each spectrum in the full biological data set.

3.2.2 Estimation of y_0

The extended glog is given in equation 3.2 (page 59) and has a second parameter, y_0, which must also be calibrated for each data set prior to use in a similar manner to the first parameter, λ. As illustrated in figure 3.1, y_0 shifts the transformation function so that the bins with the lowest intensities are scaled by the section of the glog function which has a relatively small slope. y_0 was calibrated by first estimating the noise contained within the spectra of technical replicates. The noise level was set to the smallest standard deviation of those calculated for 32 equally sized regions across the spectra [16]. The shift y_0 of the glog function was then determined by calculating the point in glog where the slope of the function increases, i.e. by calculating the point where the second derivative of z in equation 3.5 has its maximal value.

$$\frac{d^2 z}{dy^2} = -y(y^2 + \lambda)^{-\frac{3}{2}} \tag{3.5}$$

This point was typically set to three times the noise value of the spectrum. This shift ensures that the noise of the spectrum is minimally scaled by the flat region of the glog function while the larger intensity bins remain transformed using the higher values. Thus

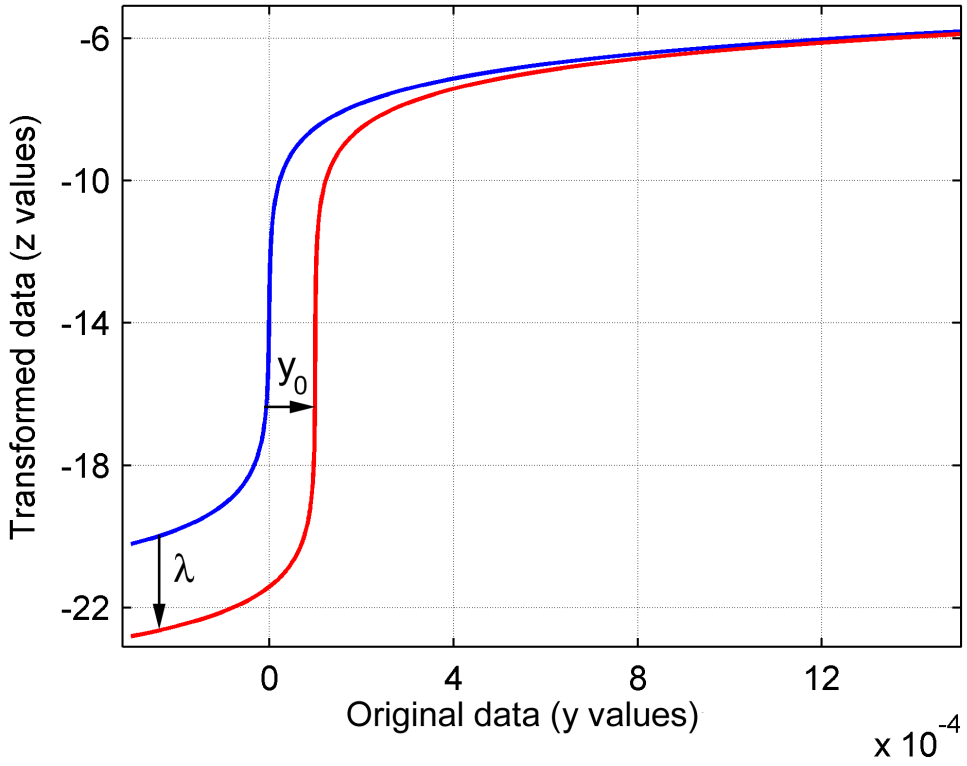


Figure 3.1: Plot of the generalised logarithm and extended generalised logarithm functions. The glog was plotted using a λ value of 1×10^{-12} (solid blue line) and the extended glog was plotted using a λ value of 1×10^{-13} and a y_0 value of 1×10^{-4} (solid red line). The effects of changing the two transformation parameters are indicated on the diagram by the black arrows. Since the transformed intensities include negative values, following the transformation each spectrum is linearly shifted upwards so that each baseline is located at zero intensity.

the noise is effectively suppressed relative to those bins corresponding to low and medium intensity peaks. Since y_0 depends on the choice of λ the optimisation of λ must be carried out first followed by the calculation of y_0. In some cases it may be necessary to optimise λ a second time after y_0 has been set, in particular for very noisy data.

Since local minimsation methods were used for both calibration methods described here, the minimisation routine was terminated when the absolute change in λ was less than a predetermined value (here 1×10^{-16}) or a maximum number of iterations was completed (here 1×10^{3}). Whilst this limits the run time of the calibration, it also increases the possibility of the optimisation failing to converge at a value. Table A.1 contains the optimised λ and y_0 values for the glog and extended glog transformations used in this thesis.

3.3 Scaling intercomparison

To evaluate the effectiveness of the three variance scaling methods, a small study was conducted using three disparate data sets. The data used was chosen specifically to address the applicability of the scaling methods discussed above in section 3.1 on typical sample types acquired in metabolomic experiments. The three data sets used were:

1. Urine samples acquired from two different dog breeds [72];
2. Muscle tissue extract taken from hypoxic and control marine mussels [22];
3. Liver tissue from fish collected from two British rivers with different amounts of pollution.

Since one goal of variance scaling methods is to aid classification, each technique is evaluated on the basis of its effects on the results of the PCA of the data. For consistency and clarity, only the first two principal components of each are used to construct the loadings plot describing the spread of the data. Linear discriminate analysis (LDA) is

then used to evaluate the clustering of samples, generating statistics of sensitivity and specificity for comparison. Effects of the algorithms upon biomarker discovery are also investigated, since investigating biomarkers is also a primary goal of many metabolomic experiments.

Each data set was acquired and processed using standard proceedures for both 1D and JRES NMR experiments (see relevant papers [72, 22, 36] for details). Each set of spectra were then scaled using each of the three methods under investigation, as well as the unscaled data. In order to aid analysis, the mean spectrum subtracted from each sample - i.e. each data set was 'mean centred' - before PCA was performed upon each set of spectra using PLS Toolbox (Eigenvector Research, Inc., Wenatchee, WA, USA). Next, using the Discriminant Analysis Toolbox (Michael Kiefte, Dalhousie University, Canada), Fisher's LDA was applied to the first and second PCs of the PCA scores plot, producing a decision region for each two-class problem. This decision region was then used to construct classification statistics of sensitivity (number of correctly classified positive results) and specificity (number of correctly classified negative results) to evaluate the effects of the scaling techniques upon each data set. Leave-one-out cross-validation was performed on the PCA-LDA models to assess the robustness of the analyses, which are shown in table B.1. The relative standard deviation (RSD, see chapter 2) for each bin, given as the standard deviation divided by the mean, was calculated for each set of technical replicates, excluding bins with an intensity lower than the estimated noise level of the spectrum (i.e., the RSD was calculated using only those bins that contained peaks). Additionally, PCA loadings plots for the 1D and pJRES data were produced by constructing the linear combination of the loadings along PC1 and PC2 that is perpendicular to the LDA decision line. The loadings plot for the 2D JRES experiment, shown in 2D matrix format to mimic an intact 2D JRES spectrum, was reconstructed from the row vector containing the loadings of the concatenated spectra. To evaluate the discriminatory potential of metabolic biomarkers discovered in the loadings plots, one-way analysis of variance (ANOVA) was

performed on each of the 5 bins with the largest absolute loadings values, for each data set and method of scaling.

3.3.1 Mussel adductor samples

Firstly, the mussel samples were analysed. These are muscle tissue dissected from the Mediterranean mussel *Mytilus galloprovincialis*, with the first group (n = 12) taken from oxygen deprived (hypoxic) animals and the second group (n = 15) from control (or normoxic) animals. An additional pooled sample was also acquired for calibrating the glog transform. The ^{1}H NMR samples were prepared and acquired as described in Hines et al [22].

1D spectra

Figure 3.2 shows the box plots of the mussel data set after each of the 4 different variance scaling methods have been applied. Here, it can be seen that each of the scalings have affected the spectral variation in different ways, with the glog transformed spectra now exhibiting the smallest median RSD, and the autoscaled data now has the fewest outliers. However, this does not help quantify the effects of the different scaling methods have upon classification. PCA was then performed and the results shown in figure 3.3. Here it can be seen that the first two principal components of the unscaled data (shown in part A) do not help separate the samples, with both classes spread throughout the plot. The LDA decision boundary correctly classifies 16 of the 27 samples, providing a benchmark to which the variance scaling methods can be compared. The autoscaled data is shown in part B, where it describes a similar picture to the unscaled data - a mix of both classes throughout the two principal components, with the exception of a single outlier. Here there is little difference after the scaling, with 15 of the 27 samples correctly classified by the LDA decision boundary. For the Pareto scaled data (part C), again, the 17 correctly classified samples appear to be randomly distributed, rather than revealing any underlying sample structure. There is a significant difference for the glog

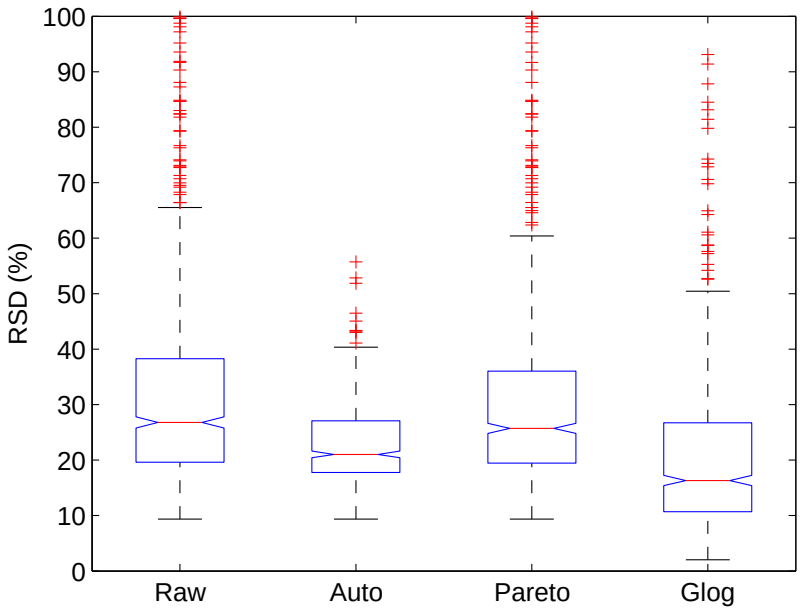


Figure 3.2: Box plots showing the percentage RSD exhibited for each of the 1D mussel adductor muscle variance scaled data sets: unscaled (raw); autoscaled; Pareto scaled and glog transformed. The notches indicate an estimate of the uncertainty about the median to facilitate box-to-box comparison. Notches with no overlap indicate a difference at the 5% significance level. Clearly, it can be seen that the autoscaled and glog transformed data have a lower median RSD than the unscaled and Pareto scaled data. The RSD range has been clipped to 0-100% to aid comparison between data sets in this thesis.

transformed spectra in part D, however. Here the samples are clearly separated, with the hypoxic samples exhibiting a positive score on the first principal component and the normoxic samples showing a negative score. The LDA decision boundary confirms this, yielding all 27 samples correctly classified. These results are summarised in table 3.1.

Figure 3.4 shows the corresponding loadings plots of the unscaled and scaled data perpendicular to the LDA decision lines shown in figure 3.3, showing the variables that best discriminate between each of the two classes. Clearly, the effects of the different scaling methods have altered the loadings plots immensely. Since the absolute size of the peaks on the loadings plots dictates how relevant that data point is to the principal component, bins with large values; either positive or negative; have the potential to be indicators of class differences. These bins can correctly identify metabolites which are

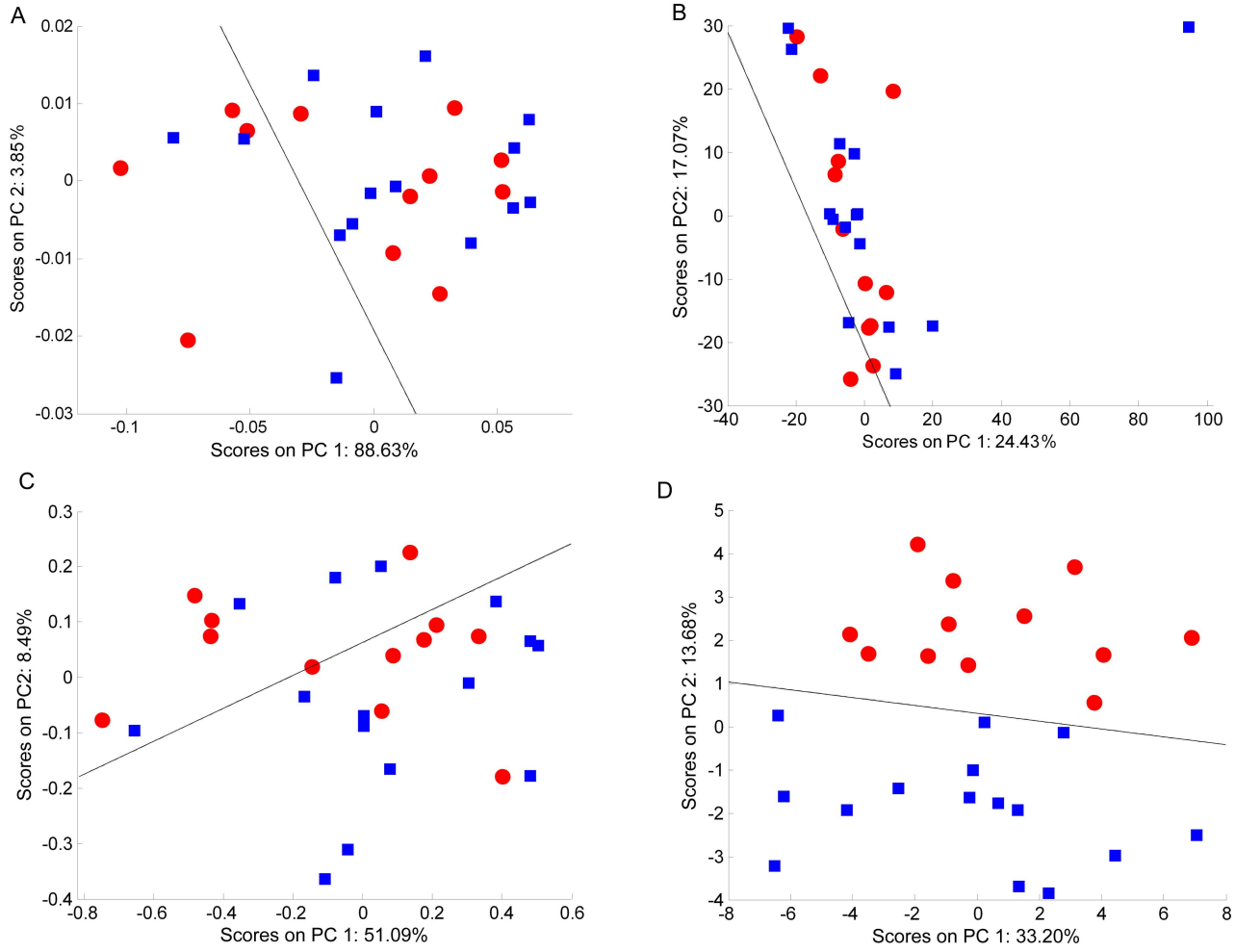


Figure 3.3: PCA scores plots of the 1D NMR spectra of mussel adductor muscle. (A) Unscaled data, (B) autoscaled data, (C) Pareto scaled data, (D) glog transformed data. The red circles represent the hypoxic samples whilst the blue squares represent the normoxic samples. The black line represents the decision boundary between the classes constructed using LDA.

Experiment & scaling	Sensitivity	Specificity	Correctly classified
1D: No scaling	0.333	0.800	16 of 27
1D: Autoscaled	0.083	0.933	15 of 27
1D: Pareto	0.500	0.733	17 of 27
1D: Glog	1.000	1.000	27 of 27

Table 3.1: Classification statistics for each PCA model constructed from the mussel adductor muscle samples.

biomarkers. To test the applicability of these potential biomarkers found by the different scaling techniques, the bins with the top 5 (absolute) loadings values from each data set were chosen and tested using one way ANOVA to determine if the bin discriminated between sample classes. Since only the glog scaled data produced a classifier with suitable accuracy, it is expected that only this data set will produce accurate biomarkers based upon the classifiers. Figure 3.4 shows that each of the bins chosen by the different scaling methods as the best biomarkers are all different. All potential biomarkers were found not to be significantly different for the unscaled, autoscaled and Pareto scaled spectra. However, the glog transformed data highlighted four highly significant bins ($p < 0.001$) and one significant bin ($p < 0.05$). These results are summarised in table 3.2.

In summary, for the mussel adductor muscle samples, the generalised logarithm performed the best, both improving the results of the PCA and by locating potentially significant biomarkers.

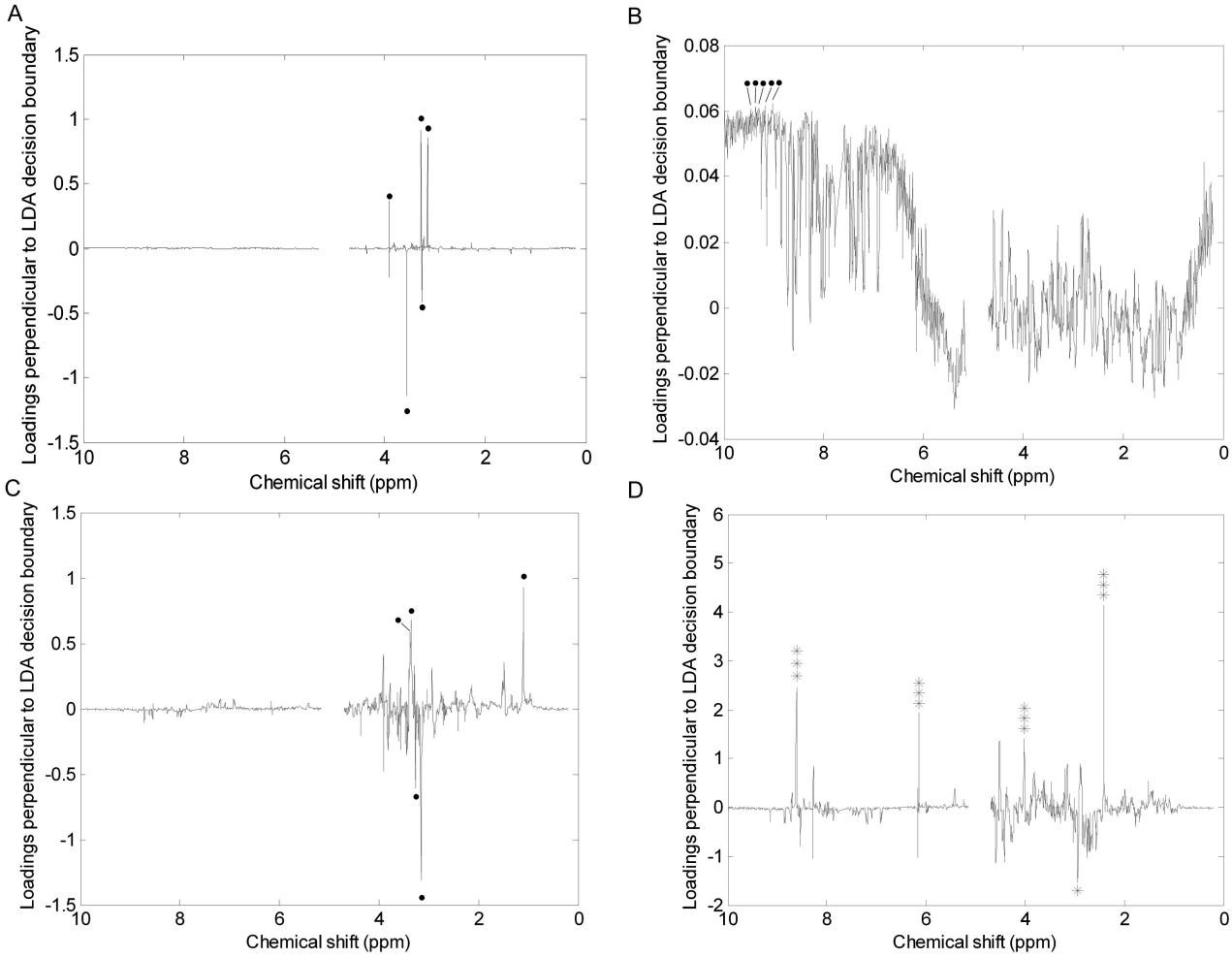


Figure 3.4: PCA loadings plots of the 1D NMR spectra of mussel adductor muscle. (A) Unscaled data, (B) autoscaled data, (C) Pareto scaled data, (D) glog transformed data. The plots represent the loadings perpendicular to the decision line calculated by using LDA on each of the scaled data sets. The 5 largest bins in each plot have each been tested as potential biomarkers to discriminate between the two classes. Key: (•) bin is not significantly different; $(*)$ $p < 0.05$; $(**)$ $p < 0.01$; $(***)$ $p < 0.001$.

Experiment	Not significant ($p \geq 0.05$)	Significant ($p < 0.05$)	Highly significant ($p < 0.01$)	Very highly significant ($p < 0.001$)
1D: No scaling	5	0	0	0
1D: Autoscaled	5	0	0	0
1D: Pareto	5	0	0	0
1D: Glog	0	1	0	4

Table 3.2: Significance of potential biomarkers for the mussel data set. Here, the 5 bins with the largest (magnitude) loadings from each scaling method are tested by one-way ANOVA to evaluate their potential as biomarkers to discriminate between classes.

3.3.2 Canine urine samples

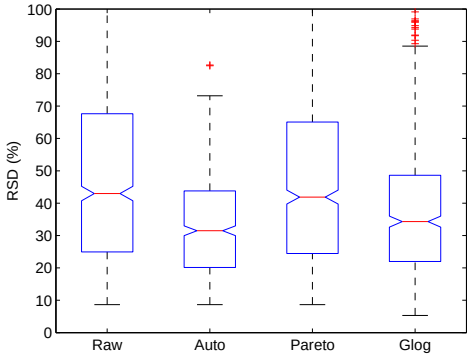
Urine samples from two different breeds of dogs - 17 samples from three male Labradors and 20 samples from four male Miniature Schnauzers - were collected and processed as described in Viant et al [72]. 1D ^1H, pJRES and JRES spectra were acquired in order to help determine the effectiveness of the scaling methods over a broad selection of both samples and experiment types..

RSD of all experiment types

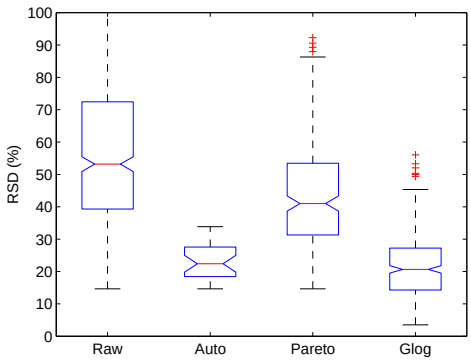
Figure 3.5(a) compares the RSD obtained from the 1D canine urine spectra after each of the different scaling algorithms have been applied. The autoscaled and glog transformed data exhibit smaller median RSD values than the unscaled and Pareto scaled data in a very similar manner to that observed for the 1D mussel adductor muscle samples (figure 3.2). Figure 3.5(b) shows the the same information derived from the pJRES spectra, where the effects of the variance scaling is more dramatic than for either of the 1D spectra data sets, as each of the scaled data sets has a much smaller median RSD value than the unscaled data. It is also clear that the glog transformed and autoscaled data sets produced the smallest RSD ranges. The data for the intact 2D JRES spectra (figure 3.5(c)) again repeats this pattern of the scaling reducing the range and median RSD values, indicating that whilst all of the scaling methods reduce the variance in the data, each scaling changes the data in a different manner.

1D spectra

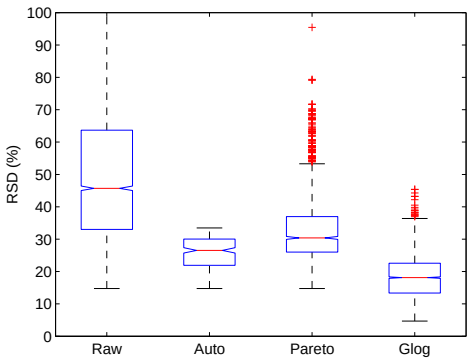
After applying PCA to the 1D spectra, the unscaled canine urine data shown in figure 3.6A produces a much better classifier than for the case of the unscaled mussel data, correctly classifying 35 of the 37 samples. Each of the scaled data sets produce slightly better classifiers, each mis-classifying only a single Labrador sample. This accuracy is reflected in the high probability of the bins indicated by the largest loadings of the analysis of each of the classifiers yielding potential biomarkers (shown in table 3.4, page 77).



(a) 1D spectra



(b) pJRES spectra



(c) JRES spectra

Figure 3.5: Box plots showing the percentage RSD exhibited for each of the (a) 1D, (b) pJRES and (c) JRES dog urine variance scaled data sets. The notches indicate an estimate of the uncertainty about the median to facilitate box-to-box comparison. Notches with no overlap indicate a difference at the 5% significance level. Clearly, it can be seen that the autoscaled and glog transformed data have a lower median RSD than the raw (unscaled) and Pareto scaled data for all experiment types. The plots have been clipped to show the 1-100% RSD range to aid inter-experiment comparision.

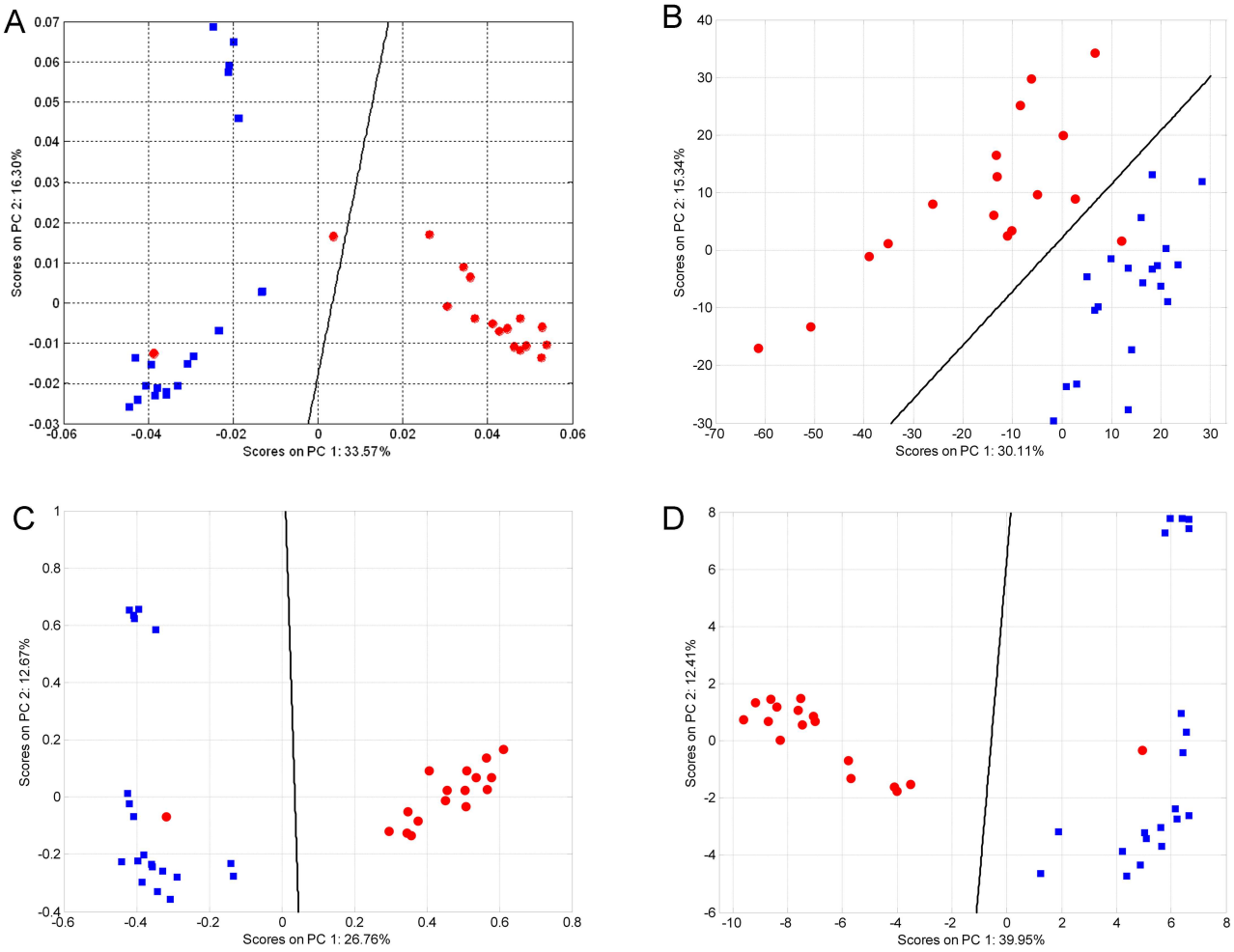


Figure 3.6: PCA scores plots of the 1D spectra of the canine urine samples. (A) Unscaled data, (B) autoscaled data, (C) Pareto scaled data, (D) glog transformed data. The red circles represent the samples from Labradors, with the blue squares representing the Miniature Schnauzer samples. The black line represents the decision boundary between the classes constructed using LDA.

pJRES spectra

The PCA scores of the pJRES spectra shown in figure 3.7 depicts a different situation to the 1D data. Here the initial classification based on the unscaled spectra is poor, as well as that based on the Pareto scaled data. However, the autoscaled and glog transformed data both produce more accurate classifiers, each misclassifying the same six samples. Examining the loadings plots for potential biomarkers (table 3.4) reinforces the good results of the autoscaled and glog transformed data, as all 5 bins representing the top loadings are classified as highly significant between classes; therefore have a high chance of being biomarkers for differentiating the two species of dogs. Both the unscaled and Pareto scaled data no not identify any potential biomarkers and so fail to reveal any useful information.

JRES spectra

Finally, the 2D JRES spectra can be examined and is shown in figure 3.8 and table 3.3. Here, the effects of the autoscaling and Pareto scaling have improved the classification rate of the PCA-LDA when compared with the unscaled data, however only 24 and 29 of 38 samples, respectively, have been correctly identified, yet the spread of the data seems random. The glog transformed data shows more structure than the other data sets, with a clear gap between two clusters of data. However, only 31 of 38 samples are correctly classified and it is necessary to check for the presence of potential biomarkers in order to assess the scaling methods further. The target loadings were used to choose bins whose potential as biomarkers was to be assessed and the results are listed in table 3.4. Only the analysis of the glog transformed data discovered bins that were significantly different, which raises the possibility of these bins being used as biomarkers; the other tested bins that were not significantly different between the two classes.

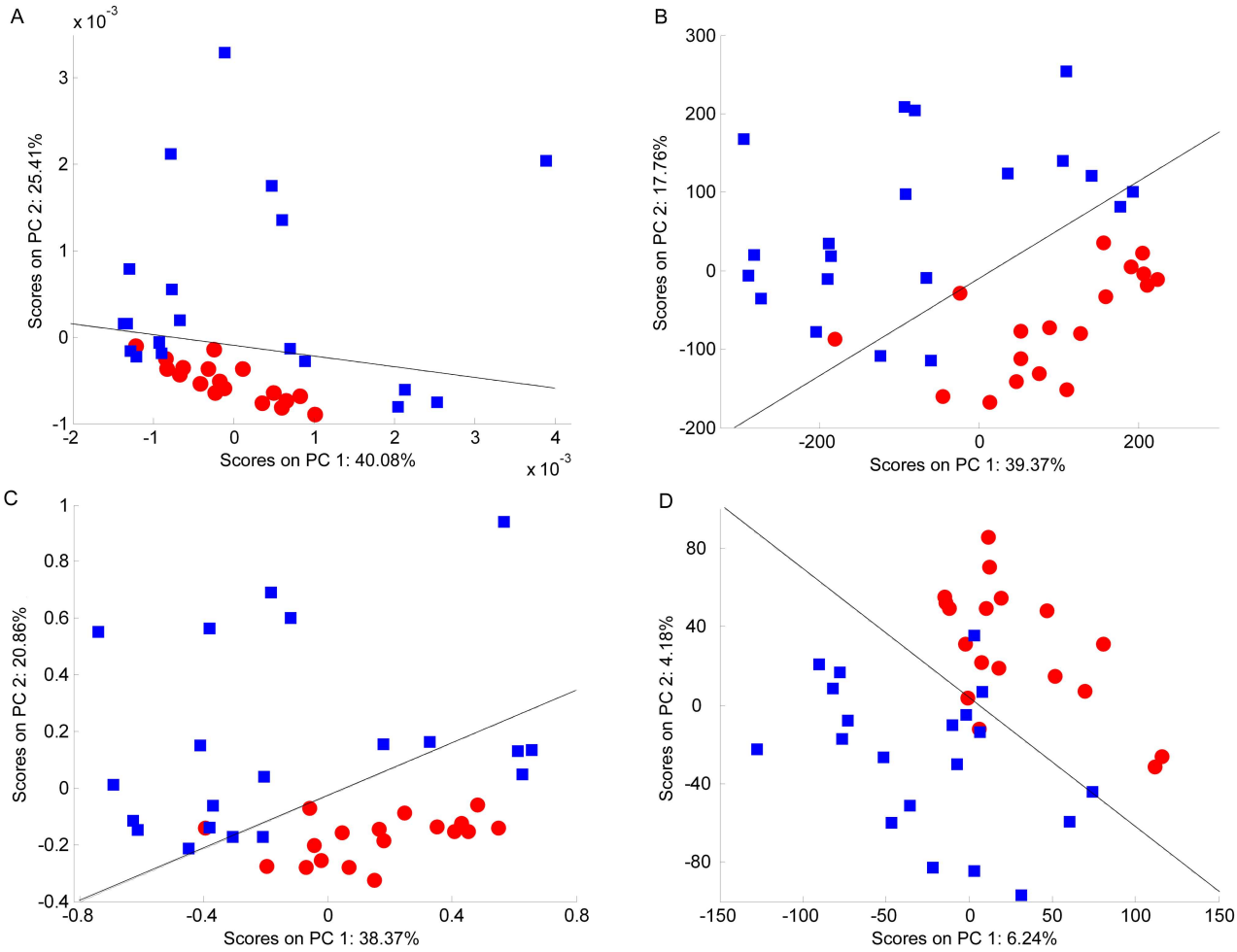


Figure 3.7: PCA scores plots of the pJRES spectra of the canine urine samples. (A) Unscaled data, (B) autoscaled data, (C) Pareto scaled data, (D) glog transformed data. The red circles represent the samples from Labradors, with the blue squares representing the Miniature Schnauzer samples. The black line represents the decision boundary between the classes constructed using LDA.

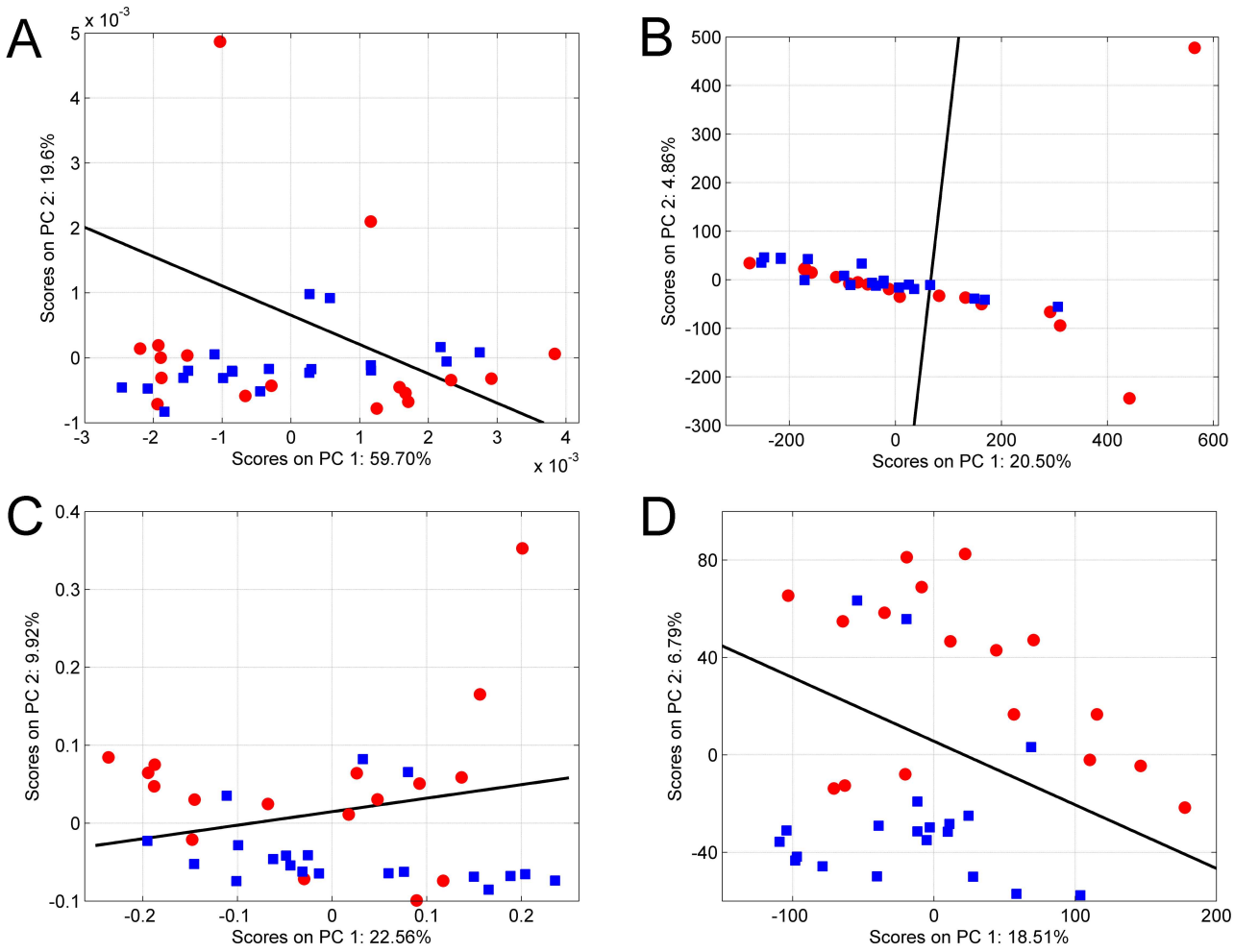


Figure 3.8: PCA scores plots of the JRES spectra of the canine urine samples. (A) Unscaled data, (B) autoscaled data, (C) Pareto scaled data, (D) glog transformed data. The red circles represent the samples from Labradors, with the blue squares representing the Miniature Schnauzer samples. The black line represents the decision boundary between the classes constructed using LDA.

Experiment and scaling	Sensitivity	Specificity	Correctly classified
1D: No scaling	0.882	1.000	35 of 37
1D: Autoscaled	0.941	1.000	36 of 37
1D: Pareto	0.941	1.000	36 of 37
1D: Glog	0.941	1.000	36 of 37
pJRES: No scaling	0.294	0.750	20 of 37
pJRES: Autoscaled	0.824	0.850	31 of 37
pJRES: Pareto	0.530	0.700	23 of 37
pJRES: Glog	0.823	0.850	31 of 37
2D JRES: No scaling	0.294	0.750	20 of 37
2D JRES: Autoscaled	0.412	0.850	24 of 37
2D JRES: Pareto	0.706	0.850	29 of 37
2D JRES: Glog	0.823	0.850	31 of 37
2D JRES: Extended glog	0.823	0.850	31 of 37

Table 3.3: Classification statistics for each PCA model constructed from the canine urine samples.

Summary of results

Again, the glog transformation has performed consistently well for the canine urine samples; producing the best or joint best classifiers which easily identify many potential biomarkers.

Experiment	Not significant ($p \geq 0.05$)	Significant ($p < 0.05$)	Highly significant ($p < 0.01$)	Very highly significant ($p < 0.001$)
1D: No scaling	0	3	0	2
1D: Autoscaled	0	0	0	5
1D: Pareto	0	0	0	5
1D: Glog	0	2	0	3
pJRES: No scaling	5	0	0	0
pJRES: Autoscaled	0	0	0	5
pJRES: Pareto	5	0	0	0
pJRES: Glog	0	0	0	5
2D JRES: No scaling	5	0	0	0
2D JRES: Autoscaled	5	0	0	0
2D JRES: Pareto	5	0	0	0
2D JRES: Glog	0	0	0	5
2D JRES: Extended glog	0	0	0	5

Table 3.4: Significance of potential biomarkers for the dog urine data set. Here, the 5 bins with the largest (magnitude) loadings from each scaling method are tested by one-way ANOVA to evaluate their potential as biomarkers to discriminate between classes.

3.3.3 Flounder liver samples

The European flounder liver samples were the last data set to be analysed. Here the two classes consist of fish sampled from two UK rivers to investigate the effects of pollution on the metabolomic profile of the fish. The two rivers were the Alde (clean/control) and the Tyne (polluted). The collection procedures are as described above and elsewhere [36] and again, both 1D ^1H spectra and 2D JRES spectra where acquired.

1D spectra

For the 1D data, little can be determined about the impact of the differences between the two classes, particularly as there is little improvement in classification between the differently processed data sets. In particular, for each of the scores plots (shown in figure 3.9) the two classes are loosely grouped, but the clusters overlap, resulting in five misclassified samples. It can be seen that each of the processing methods have resulted in tighter clustering than the unscaled data, which improves the classification rate. In particular, the glog transformed data has the best rate of classification where the samples from the River Alde are spaced closely together.

pJRES spectra

For the pJRES of the flounder liver samples, no single scaling method produces a perfectly accurate classifier. However, both the Pareto scaling and glog transformation produce better classifiers than the raw data as shown in table 3.5 and figure 3.10. Unusually, the autoscaled data reduces the accuracy of the classifier in this case. It can be seen from table 3.6 that each of the data sets fail to indicate very highly significant ($p< 0.001$) potential biomarkers, however highly significant ($p< 0.01$) bins were indicated in each data set apart from the autoscaled samples. These results indicate that the pJRES experiment type does not discriminate as well between the two sample types as the 1D experiments, as both the classifier accuracy and potential biomarker discovery results for the 2D projections were inferior.

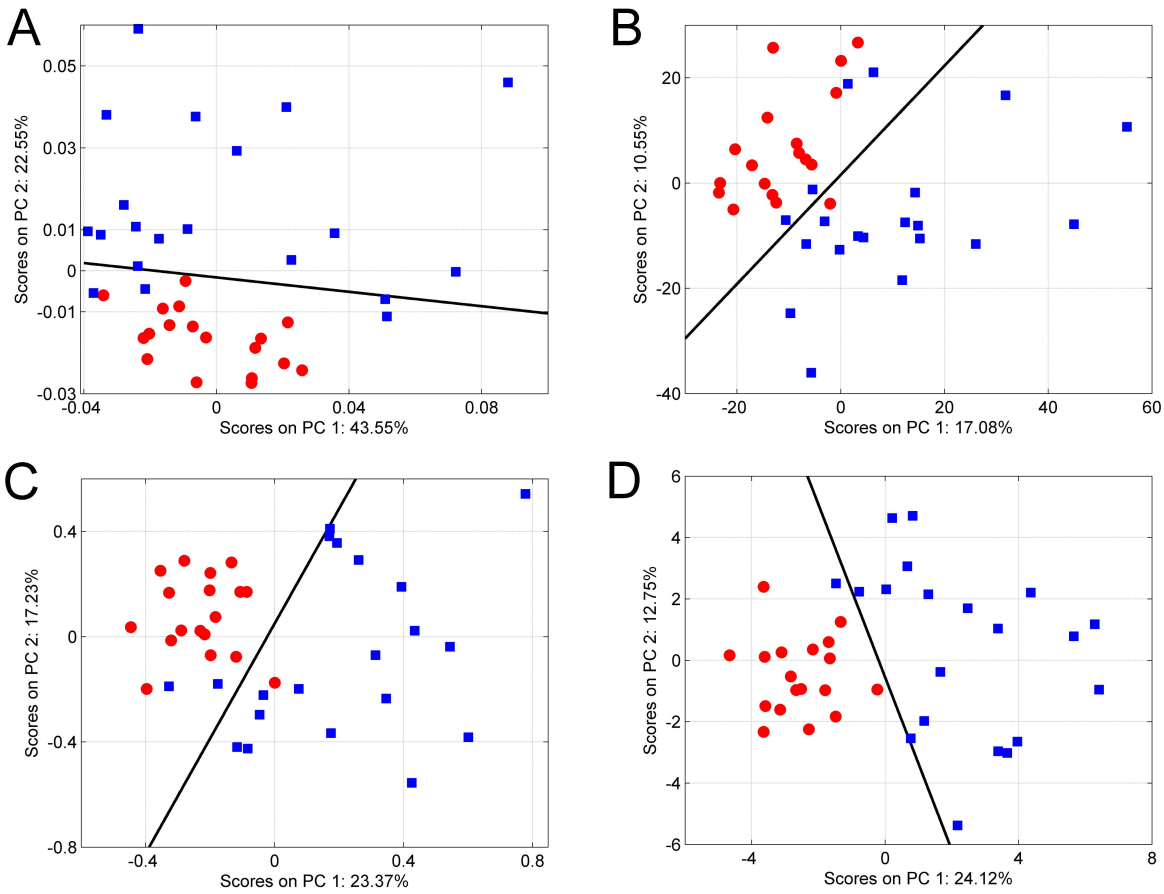


Figure 3.9: PCA scores plots of the 1D spectra of the flounder liver samples. (A) Unscaled data, (B) autoscaled data, (C) Pareto scaled data, (D) glog transformed data. The red circles represent the samples from River Alde, with the blue squares representing the River Tyne samples. The black line represents the decision boundary between the classes constructed using LDA.

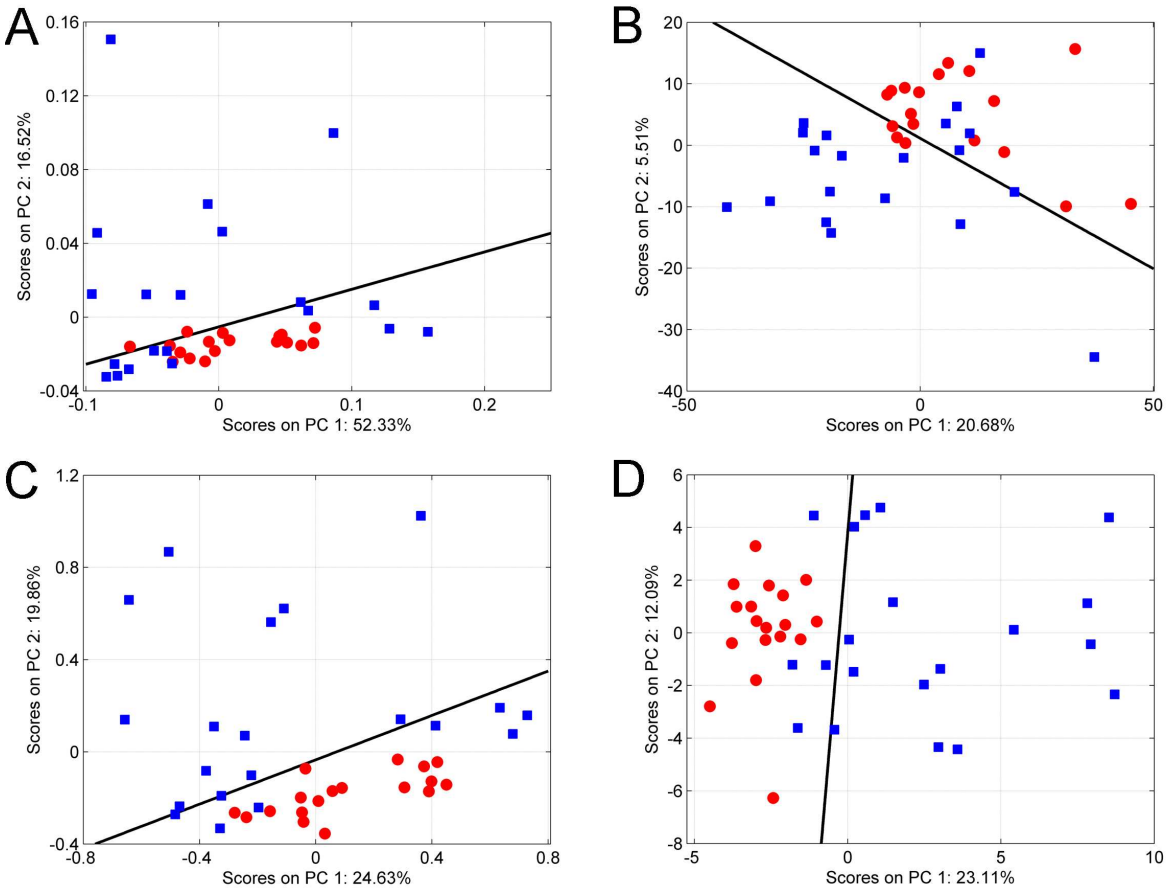


Figure 3.10: PCA scores plots of the pJRES spectra of the Flounder liver samples. (A) Unscaled data, (B) autoscaled data, (C) Pareto scaled data, (D) glog transformed data. The red circles represent fish sampled from the River Alde and the blue squares represent fish from the River Tyne. The black line represents the decision boundary between the classes constructed using LDA.

JRES spectra

Each of the PCA scores plots generated by the scaled JRES data sets produces an imperfect classifier. As shown in figure 3.11, in all cases there is partial separation of the two classes and the LDA decision line reveals that 29 of 38 samples are correctly classified for the unscaled data. This then improves to 33 of 38 samples following autoscaling, Pareto scaling or glog transformation (table 3.5). Further examination of the glog transformed data set reveals that only a small proportion of the variance is captured by the first two PCs in the PCA model. This unanticipated result can be explained by examining the glog transformed data itself. Figure 3.12A shows a glog transformed 2D JRES spectrum following concatenation of each of the slices along the J-coupling axis into a single row vector. The effect of the glog transformation (figure 3.12B) has not only increased the heights of the small peaks relative to the larger ones, but has also greatly magnified the noise in the spectrum. Using the extended glog, however, limits the effects upon the small noise peaks in the spectrum. This effect can be seen in figure 3.12, which shows the same 2D JRES spectrum after application of the ex-glog transformation. The PCA scores plot of the ex-glog transformed data is also changed and is shown in figure 3.13A where the variance expressed by the first two PCs almost doubled (compared with the standard glog) to 12.1% and 6.9%, respectively. Using the extended glog transform also improves the LDA classifier, with all 38 of 38 samples now correctly assigned to their correct classes and separation between the two classes in PCA space now readily apparent - a vast improvement over all other scaling methods.

The corresponding PC1 loadings plot for the scores plot in 3.13A is shown in two different orientations in figure 3.13 part B (top view) and part C (side view). When used in combination with the extended glog transformation the resulting loadings plot provides a powerful visualisation tool from which the metabolic differences between the two sample classes can be identified, since the preservation of J coupling information and reduction in spectral congestion can aid both peak identification and potentially, quantification.

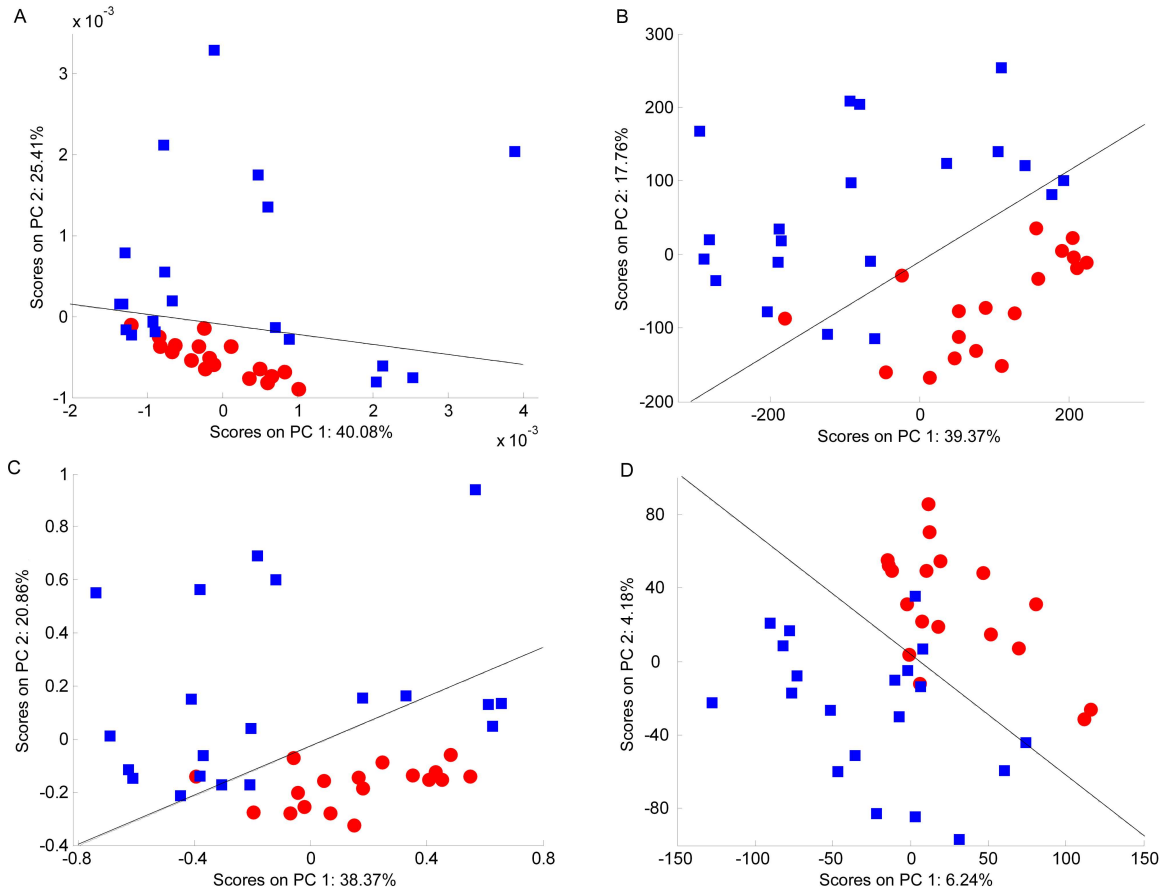


Figure 3.11: PCA scores plots of the JRES spectra of the Flounder liver samples. (A) Unscaled data, (B) autoscaled data, (C) Pareto scaled data, (D) glog transformed data. The red circles represent fish sampled from the River Alde and the blue squares represent fish from the River Tyne. The black line represents the decision boundary between the classes constructed using LDA.

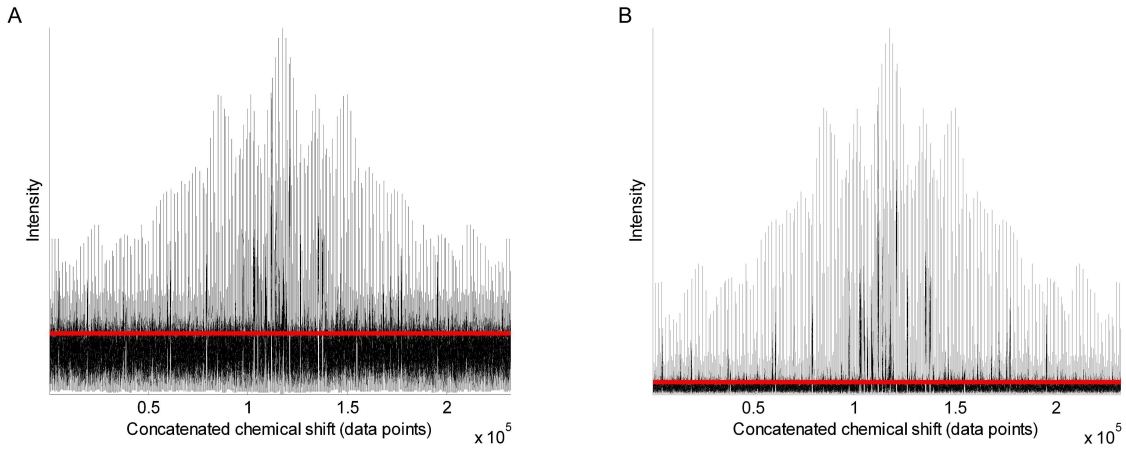


Figure 3.12: The concatinated profile of the European flounder liver data after the glog (part A) and extended glog (part B) transformation. The noise level is indicated by the solid line. The extended glog transform has a much smaller noise component

Summary of results

Despite application of each of the scaling methods, the flounder liver samples remained difficult to separate using 1D and pJRES spectra, as all three initial scaling methods failed to produce a perfect classifier. The use of intact 2D spectra did not initially aid classification, until steps were taken to reduce the large impact that the glog transformation has upon the noise within the spectra. With the addition of an extra parameter, the extended glog transformed data then produced two clearly separable clusters when using PCA-LDA.

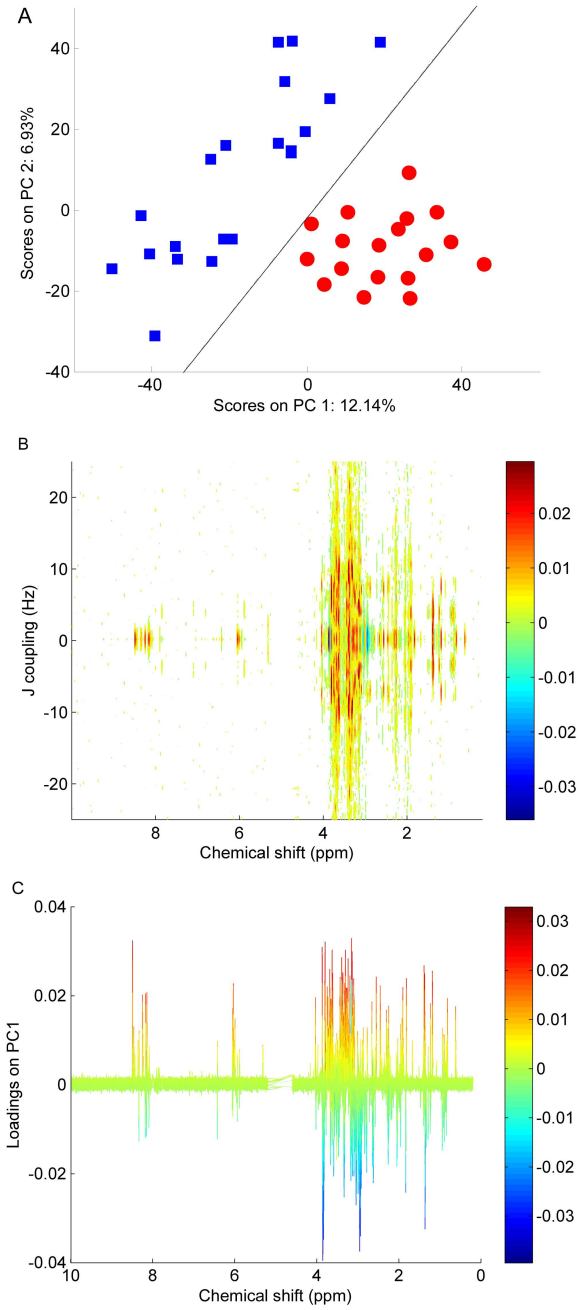


Figure 3.13: PCA plots of the extended glog transformed JRES spectra of the European flounder liver samples. (A) Scores plot where red circles represent the fish sampled from the River Alde and the blue squares represent fish from the River Tyne. The black line represents the decision boundary between the classes constructed using LDA. (B) Aerial view of the corresponding PC1 loadings plot presented in the format of a 2D JRES spectrum, with J-couplings along one axis to facilitate metabolite identification. (C) Side view of the same loadings plot as in B, highlighting the metabolites that are at higher concentration (red) in fish liver collected from the River Alde.

Experiment and scaling	Sensitivity	Specificity	Correctly classified
1D: No scaling	0.889	0.450	25 of 38
1D: Autoscaled	0.833	0.750	30 of 38
1D: Pareto	1.000	0.650	31 of 38
1D: Glog	1.000	0.800	34 of 38
pJRES: No scaling	1.000	0.550	29 of 38
pJRES: Autoscaled	0.556	0.800	26 of 38
pJRES: Pareto	0.944	0.900	35 of 38
pJRES: Glog	0.889	0.850	33 of 38
2D JRES: No scaling	1.000	0.550	29 of 38
2D JRES: Autoscaled	0.944	0.800	33 of 38
2D JRES: Pareto	0.944	0.800	33 of 38
2D JRES: Glog	0.889	0.850	33 of 38
2D JRES: Extended glog	1.000	1.000	38 of 38

Table 3.5: Classification statistics for each PCA model constructed from the flounder liver samples.

Experiment	Not significant ($p \geq 0.05$)	Significant ($p < 0.05$)	Highly significant ($p < 0.01$)	Very highly significant ($p < 0.001$)
1D: No scaling	0	2	2	1
1D: Autoscaled	0	2	1	2
1D: Pareto	0	1	2	2
1D: Glog	0	1	2	2
pJRES: No scaling	1	0	4	0
pJRES: Autoscaled	2	3	0	0
pJRES: Pareto	1	0	4	0
pJRES: Glog	1	0	4	0
2D JRES: No scaling	1	0	4	0
2D JRES: Autoscaled	2	3	0	0
2D JRES: Pareto	0	0	5	0
2D JRES: Glog	0	5	0	0
2D JRES: Extended glog	0	0	0	5

Table 3.6: Significance of potential biomarkers for the European flounder liver data set. Here, the 5 bins with the largest (magnitude) loadings from each scaling method are tested by one-way ANOVA to evaluate their potential as biomarkers to discriminate between classes.

3.3.4 Summary of comparisons

In this chapter it has been demonstrated that autoscaling, Pareto scaling and the glog and extended glog transformations can significantly alter the variance structure of NMR metabolomics data, which in turn can improve the classification accuracy of multivariate models generated from the scaled data. This can help to extract information from data sets, since improving the discrimination between sample classes can help to identify metabolic biomarkers. Specifically, it has been demonstrated that the glog and extended glog transformations achieve the best, or equal best, classification accuracy compared with unscaled, autoscaled and Pareto scaled data on three example data sets. A high classification accuracy was achieved for the majority of the data sets using these scaling methods. Furthermore, from an analysis of the top five peaks in each of the corresponding PCA loadings plots, it has been shown that the glog transformed data demonstrates considerable accuracy at discovering spectral bins yielding metabolic biomarkers that can discriminate significantly between sample classes. This study was limited to the use of the first two principal components only (to ensure a fair comparison), which may not always produce the best axis to separate the data. Restricting the study to using solely PCA may also limit the results, since PCA is not the best tool to use for non-linearly separated data. However, this study does show that the choice of variance scaling transformations can drastically alter the results of the analysis. The broad applicability of the glog approach using three disparate data sets from different biological samples has also been shown; using 1D, pJRES and intact 2D JRES spectra. Finally, we have reported an extension to the original glog algorithm that improves the signal to noise ratio in the transformed spectra, which was critical for the accurate analysis of intact 2D JRES spectra of the European flounder samples. In conclusion, the scaling methods described in this section have been thoroughly evaluated and their benefits to data processing highlighted. In particular, the advantages of utilising the glog transformation for stabilising the technical variance associated with metabolomics experiments has been shown: this can lead to significantly beneficial effects upon classification of samples using multivariate

analysis.

Chapter 4

JRES NMR

Due to the severe spectral overlap of resonances in 1D ^{1}H NMR spectra of biological samples, unambiguous identification and quantification of individual metabolites is difficult. Knowing which metabolites are present in a complicated biological sample is highly desirable and consequently there is an increasing use of 2D NMR experiments, as these spread the resonances into a second dimension [36, 71, 73, 75]. As discussed in section 3.3, the ^{1}H J-resolved (JRES) experiment is one of the most popular 2D methods used in metabolomics[36, 71, 73, 75] as this approach can provide a less congested metabolic fingerprint than 1D NMR, yet in a relatively short acquisition time compared with most 2D NMR experiments due to the low number of increments recorded in the indirect dimension. To date, the typical processing of JRES metabolomics data has comprised of taking a 1D projection of each 2D spectrum, providing a simple format for multivariate statistical analysis[71]. However, calculation of a 1D projection discards potentially important data, the spin-spin coupling pattern, which could be used to further discriminate between different metabolites within a complex biological sample. In fact, spin-spin coupling measurements could prove of significant benefit for metabolite identification since they are less sensitive to changes in pH than chemical shift values[43]. This chapter presnts an investigation into the use of intact 2D JRES spectra for use in metabolomics experiments.

In this chapter, the processing methods specific to the JRES experiment are discussed. Specifically, the effects of the type of apodization function used to process the data and the consequences upon both pJRES and 2D JRES spectra. The spectral resolution of 2D JRES data is also discussed. Finally, a summary of the findings and recommendations for JRES NMR are made.

4.1 Window functions and projection methods

In previous NMR investigations of biological samples where JRES spectra have been used, data processing was performed using either a sine-bell [12, 36, 71, 73, 79] or squared sine-bell [13, 42] window function prior to Fourier transformation of the free induction decay (FID) to obtain the 2D spectrum. Subsequently, a summation [2] or skyline [12, 71, 73] projection method was applied to obtain the 1D projected spectrum, as outlined in section 1.3.4. Here, the traditional sine-bell window function is compared with the combination of the sine-bell and the exponential window functions. The data processing methods are evaluated in terms of maximum signal to noise ratios, spectral reproducibility and resolution and for the presence of baseline artifacts, both for pJRES and intact 2D JRES spectra. The spectral quality of the 1D pJRES data obtained using skyline and summation projection methods is also presented. The complete investigation (also involving 1D spectra) has been published by Analyica Chimica Acta by Tiziani et al [65][1].

4.1.1 Data description

To confirm the widespread applicability of the optimised processing scheme, the studies use NMR spectra from three disparate, but commonly used types of biological samples: mammalian urine (dog); fish liver extract (European flounder) and cell extract (K562 acute myeloid leukaemia cell line). Five technical replicates were created from each type of biological sample and, in turn, each of the sample types were individually prepared and

[1]In this section, the acquisition of the data and the analysis of the pJRES spectra was performed by S. Tiziani and A. Lodi. The analysis of the intact 2D JRES spectra was performed by the author.

analysed [65].

Spectra were collected using methods described elsewhere (section 1.3.4, Tiziani et al.[65]). Specific to this investigation, however, window functions were applied prior to Fourier transformation in both dimensions. In the direct dimension either a sine-bell (SINE) or a combined sine-bell and exponential (SEM; with 0.3 Hz exponential line broadening, LB) window function was utilised. In the indirect dimension a sine-bell function was used in all cases. Following Fourier transformation, the magnitude mode spectra were tilted by 45° and symmetrised. These are standard processing steps which are introduced in section 5.3, however, the impact and necessity of 'symmetrisation' and 'tilting' is also investigated in chapter 6.

Projections of each 2D JRES spectrum were performed by both skyline and summation methods. For a skyline projection, the maximum intensity along the indirect (J coupling) dimension was taken, whilst the summation projection was obtained by summing all the values along the indirect dimension. The data were further processed using NMRLab [20] running within MATLAB, including spectral alignment, exclusion of unwanted signals, TSA normalisation and binning at 4 data points per bin (approximately 0.005 ppm). The noise level of each binned pJRES spectrum was evaluated using the method described by earlier in section 2.1.2 [16]. Finally, both the intact 2D JRES spectra and their associated noise surface matrices were concatenated into a single row vector. Spectral quality parameters were then calculated as described below.

Spectral quality parameters

Spectral sensitivity was estimated for the two different processing methods by counting the number of bins in each spectrum with signal intensity greater than three times the spectral noise level (calculated separately for each spectrum). The signal-to-noise ratio (SNR) for the most intense peak in each spectrum was also used as an indication of sensitivity. The *reproducibility* across the five technical replicates was assessed by calculating the RSD of

the signal intensity in each bin of the spectrum, whilst the *resolution* obtained from the different processing methods was evaluated based on a direct visual comparison of the non-binned NMR spectra.

4.1.2 Investigation results and recommendations

The numbers of bins containing signal for each of the pJRES processing methods for all sample types are shown in table 4.1. Here it can be seen that the SEM window function generated more signal bins than the SINE processed data for both the summation and skyline projection methods. Similarly, the skyline projections performed better than the spectra produced by summation projection. This is due to the skyline projections producing lower noise thresholds, possibly due to a noise-smoothing effect of the skyline projection since the maxima picked by this procedure in noise regions of the spectrum are typically similar in value. This decrease in noise is reflected in the higher signal to noise ratios (SNR) which are recorded in table 4.2. It can also be seen that the RSD (table 4.3) of each of the different sets of technical repeats reflects this pattern: the SEM spectra are more reproducible (i.e. have a lower median RSD) than the SINE spectra; and the skyline projections tend to be be equally or more reproducible than the summation projection. The JRES spectra also follow the trend of the pJRES spectra, with the SEM apodized data producing more signal bins, higher SNR and lower RSD.

In summary, it can then be concluded that the SEM apodization produced superior results than the SINE apodised spectra, and that pJRES spectra should be formed by taking skyline projections rather than the summation method.

	Number of bins containing signal			
	SEM skyline	SEM sum[a]	SINE skyline	SINE sum[a]
Dog urine				
Rep 1	1054	952	671	635
Rep 2	1040	965	705	655
Rep 3	1041	1000	695	605
Rep 4	1013	1013	686	657
Rep 5	1060	973	709	619
Mean	1042	981	693	634
Std deviation	18	25	15	23
No. consistent bins[b]	972	893	612	529
Fish liver				
Rep 1	587	528	404	366
Rep 2	552	514	362	333
Rep 3	577	523	386	335
Rep 4	555	526	383	349
Rep 5	565	516	368	310
Mean	567	521	381	339
Std deviation	15	6	16	21
No. consistent bins[b]	510	470	326	285
Leukaemia cell				
Rep 1	750	699	550	555
Rep 2	776	717	576	533
Rep 3	793	721	608	535
Rep 4	792	756	599	585
Rep 5	804	743	616	640
Mean	783	727	590	570
Standard deviation	21	22	27	45
No. consistent bins[b]	713	663	519	477

Table 4.1: Number of bins containing signal in the 1D pJRES spectra for all five replicates of the dog urine, fish liver and leukaemia cell extracts. [a]Projection method is summation. [b]The number of bins that contain signal in all five technical replicates.

	SEM skyline	SEM sum[a]	SINE skyline	SINE sum[a]
Peak used	Signal to noise ratio (SNR)			
Dog urine				
2.72ppm	1925 ± 197	425 ± 46	807 ± 82	169 ± 24
3.05ppm	52286 ± 4948	11618 ± 872	21294 ± 1228	4511 ± 427
3.27ppm	8264 ± 801	2212 ± 192	3295 ± 266	854 ± 62
40.5ppm	23371 ± 2173	5368 ± 480	8841 ± 546	1937 ± 212
Fish liver				
1.33ppm	731 ± 74	494 ± 18	268 ± 20	191 ± 18
1.49ppm	4049 ± 428	2239 ± 103	1509 ± 154	838 ± 80
3.04ppm	2861 ± 309	1047 ± 48	1133 ± 101	441 ± 40
3.27ppm	29781 ± 2505	10592 ± 239	9039 ± 577	3303 ± 165
Leukaemia cell				
1.33ppm	3076 ± 520	1248 ± 282	1397 ± 391	566 ± 102
3.04ppm	20366 ± 2287	4097 ± 519	9549 ± 2069	1938 ± 243
3.94ppm	8085 ± 586	1669 ± 142	3388 ± 501	711 ± 70
8.24ppm	1714 ± 179	384 ± 44	788 ± 169	161 ± 22

Table 4.2: Signal-to-noise ratios of selected metabolites in the 5 replicate spectra for each of 3 biological samples. The metabolite chosen are 1.33 ppm lactate; 1.49 ppm alanine; 2.72 ppm dimethylamine; 3.04 ppm creatine; 3.27 ppm taurine; 3.94 ppm creatine; 4.05 ppm creatinine; 8.24 ppm IMP. [a]Projection method is summation

	Dog urine	Fish liver	Leukaemia cell
1D pJRES			
SEM skyline	12.4%	11.6%	11.2%
SEM sum[a]	9.9%	7.6%	12.9%
SINE skyline	11.9%	11.9%	21.1%
SINE sum[a]	13.6%	11.4%	16.0%
2D JRES			
SEM	15.4%	16.9%	13.7%
SINE	16.3%	19.0%	17.9%

Table 4.3: Median RSD values of the binned signal intensities for projected 1D pJRES and intact 2D JRES spectra of dog urine, fish liver extract and leukaemia cell extract samples. [a]Projection method is summation.

	No. of signal bins		SNR of most intense peak	
	SEM	SINE	SEM	SINE
Dog urine (3.04 ppm)				
Rep 1	5586	2797	18660	4474
Rep 2	5194	2641	13614	5326
Rep 3	5323	2583	13870	5156
Rep 4	5235	2640	18405	5666
Rep 5	5189	2553	13276	4279
Mean	5305	2643	15565	4980
S.D.	166	94	2719	585
No. consistent bins[a]	3883	1562		
Fish liver (3.27 ppm)				
Rep 1	6934	4325	22709	8873
Rep 2	6970	4353	23278	8532
Rep 3	7180	4412	25784	8330
Rep 4	7009	4467	24348	9169
Rep 5	6935	4320	26316	9720
Mean	7006	4375	24487	8925
S.D.	102	63	1555	548
No. consistent bins[a]	4885	2654		
Leukaemia cell (3.04 ppm)				
Rep 1	4981	2746	4284	1627
Rep 2	5124	3061	5441	2529
Rep 3	5437	3117	4837	1807
Rep 4	5605	3167	5265	1906
Rep 5	5388	3141	6351	2302
Mean	5307	3046	5235	2034
S.D.	251	172	767	371
No. consistent bins[a]	4072	1940		

Table 4.4: Number of bins containing signal and the signal-to-noise ratio (SNR) of the most intense peak in the 2D JRES spectra for all five replicates of the dog urine, fish liver extract and leukaemia cell extract samples

4.2 Spectral resolution

The effects of spectral resolution have been well studied for 1D (and pJRES) experiments, where it has been found that whilst the acquision resolution of the experiment is important, so are the widths of the 'bins' of the chemical shift axis [71, 74, 31, 66].With the addition of the second frequency dimension in 2D JRES experiment, it is important to examine the effects of resolution upon both spectral and multivariate analysis.

The resolution of a JRES spectrum in the direct dimension has the same strengths and weaknesses as a 1D or pJRES spectrum. That is, by increasing the number of data points it becomes easier to identify and deconvolve peaks; unfortunately large numbers of data points increase the complexity and duration of any multivariate analysis. In this thesis, all FIDs are collected using high resolution methods, which yields spectra with typically 16,000 data points. The number of data points is then increased again by using zero filling during Fourier transformation. It is important to note that whilst zero filling improves the appearance of the spectra and had limited beneficial effects during peak picking (due to the smoothing effects), it does not add to or change the information present with the FID. The indirect dimension, however, is collected with significantly less increments (data points) than the direct dimension. For example, 8, 16 [65] or 32 [71] increments have been routinely used in metabolomics experiments. Clearly, it is necessary to select the number of increments in the indirection such that whilst peaks may be identified and deconvolved, the overall number of data points in the spectrum is low enough to allow any spectral analysis to proceed with reasonable time and ease.

4.2.1 Peak broadening

Visually examining the peak corresponding to the internal standard - in this case, TMSP - of 'high' (32 spectral increments in the indirect dimension) and 'low' (16 increments) resolution spectra highlights the first issue arising from investigating different resolutions of the indirect dimension. This is illustrated in figure 4.1, where part A shows a low

resolution peak and part B a high resolution peak. Here, both spectra have been zero filled to 128 points to ensure smooth peaks. Clearly, the two peaks are of very different widths; with the low resolution peak around 30 Hz wide and the high resolution peak is approximately 10 Hz wide. Minimally, this makes comparison between spectra collected under different conditions difficult and could, potentially, obscure peak structures such as doublets and inhibit attempts at metabolite identification and quantification.

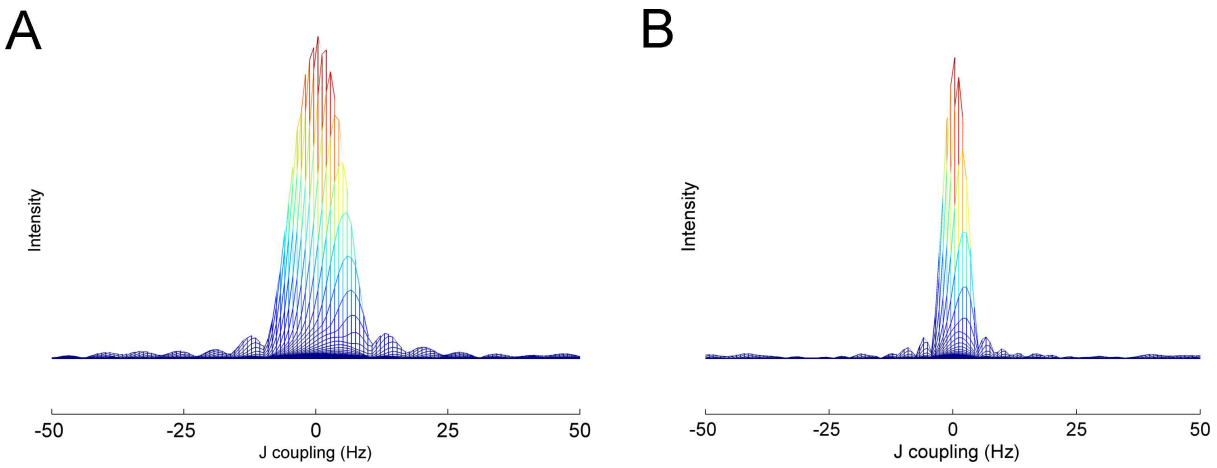


Figure 4.1: Peak widths are dependent upon the resolution of the spectrum. Here, projections of the internal standard (TMSP) peak in a low resolution (16 increments, part A) and a high resolution (32 increments, part B) spectra are shown. Clearly, the low resolution spectrum produces broader peaks. Both spectra have been zero filled to 128 points.

4.2.2 Peak fragmentation

Removing the zero filling does not limit the peak broadening, but creates other problems. Inspection of the doublet present in alanine at 1.48 ppm (figure 4.2) reveals that due to the low number of points present in the non-zero filled spectrum, the peaks begin to split and fragment as the doublet in figure 4.2B is reduced to an apparent quartet in figure 4.2A. This particular effect is only observed in tilted, yet unfolded spectra.

Folding ensures that the spectrum is symmetrical about the line of 0 Hz, so the two 'extra' peaks seen in figure 4.2A will be removed after folding has taken place. However, as figure 4.3 illustrates, due to the 'fragmentation', the shape of the peak in the low resolution

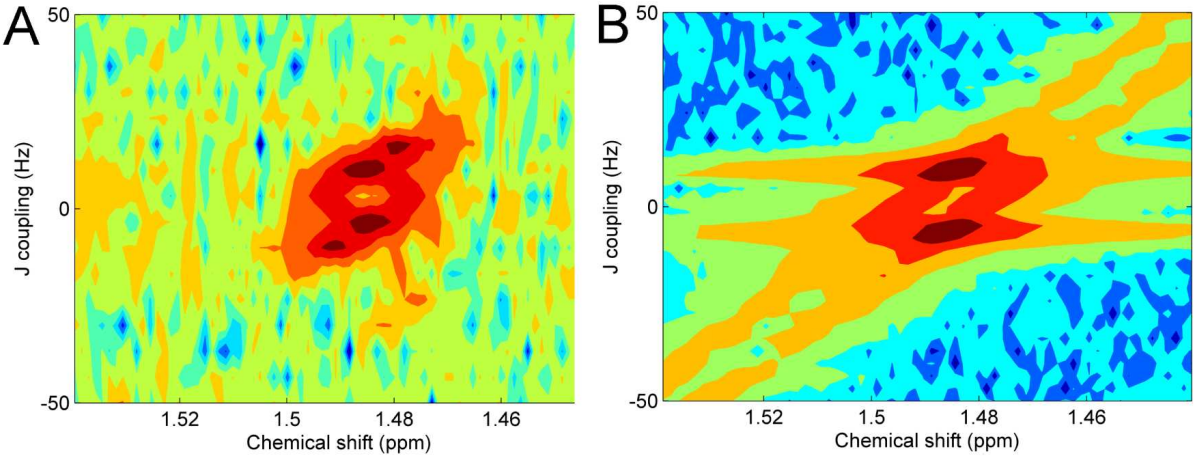


Figure 4.2: Peak fragmentation is illustrated using the alanine doublet. At low resolution (16 increments, part A), the peaks begin to fragment, which is not present in the high resolution (32 increments, part B). Both spectra have been tilted, but not folded and are shown as $\log_{10}$ intensity plots. No zero filling has been applied.

spectrum (figure 4.3A) produced after folding still exhibits some peak fragmentation. A human observer may easily discriminate between the different maxima to discern the 'true' (i.e. expected) peaks, yet this is often a difficult task to automate. It can also be seen that the two 'true' peaks are not distinct and are joined above the noise threshold, which further distorts the line-shapes of the metabolite peaks, which may complicate any peak deconvolution efforts. Whilst the higher resolution peaks (figure 4.3 B) exhibit some overlap and interference, it is greatly reduced due to the smaller frequency range in the indirect dimension that each peak encompasses. Although some interference has occurred between the two known resonances, no peak fragmentation is present.

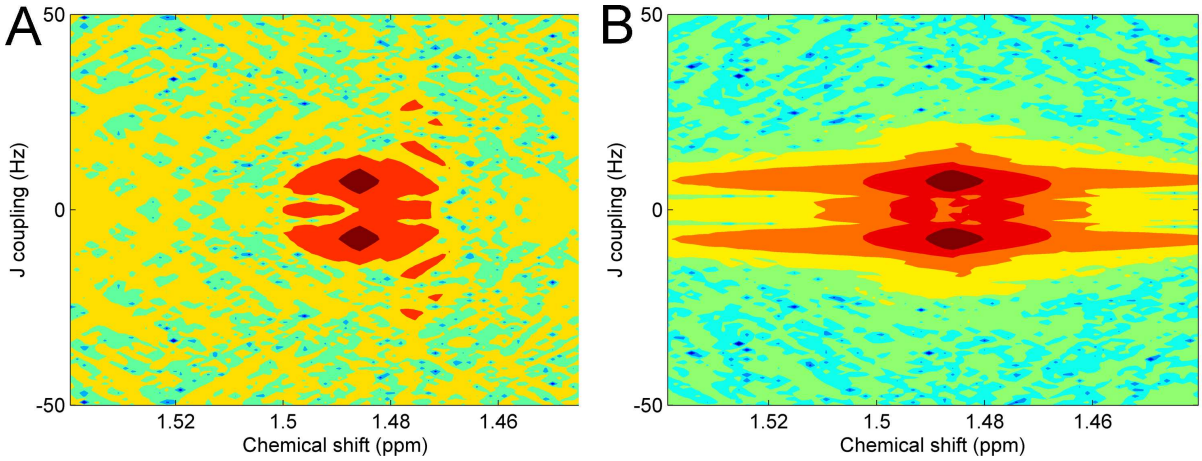


Figure 4.3: Peak fragmentation after folding is illustrated using the same alanine doublet as figure 4.2. At low resolution (16 increments, part A), the peaks have fragmented, which has spread and obscured the peak shapes, particularly at the base of the peaks (light red contour). The high resolution spectra (32 increments, part B) produce a structure more clearly recognisable as two peaks. Both spectra have been tilted, folded, zero filled to 128 points and are shown as $\log_{10}$ intensity plots.

4.2.3 Recommendations for spectral resolution

Clearly because of differences in peak widths it is necessary to compare spectra acquired and processed using parameters that are as similar as possible in order to make accurate comparisons of JRES spectra. This ensures that peaks occur with minimal shifting and intensity differences between spectra. It is also recommended to use spectra of as high as resolution as possible - minimally 32 increments in the indirect dimension. This will minimise peak fragmentation and distortion.

4.3 Summary of optimal JRES processing

This chapter has discussed the optimisation of JRES processing, from the initial acquisition parameters used for all JRES experiments to the choice of methods used to construct the pJRES spectra. By correct processing of the spectra, it ensures that any following analysis has the maximum opportunity to extract useful and meaningful information. Further, examining the projection methods used to create the pJRES spectra, it was found that the skyline projections often created more robust spectra than the summation method. By considering the effects of two different apodization functions, increases in the the signal to noise ratio were observed in both the pJRES and 2D JRES spectra when using the SEM window function. Finally, when examining the JRES specific processing methods of tilting and symmetrisation, it was noted that using 2D JRES spectra with less than 32 increments can result in peak broadening and fragmentation - two phenomena that complicate spectral peak picking and deconvolution.

It is thus recommended that all JRES spectra are acquired using high resolution methods in both the direct and indirect dimensions, whilst using the SEM apodization function. It is also recommended that pJRES spectra should be formed by taking skyline projections rather than the summation method.

Chapter 5

Analytical line shapes of JRES peaks

A number of strategies exists for the deconvolution of mixture spectra into their component parts [24, 25, 27, 34, 44, 52, 82]. One such approach - which is potentially accurate, though computationally expensive - is fitting each individual resonance in an intact 2D JRES spectrum to the theoretical line-shape for this type of NMR data. Such an approach has been employed successfully for 1D NMR spectra [1]. However, unlike for a 1D spectrum in which, under ideal conditions, each resonance is known to have a Lorentzian line-shape [4], the mathematical function describing a resonance in a processed 2D JRES spectrum has not been thoroughly evaluated [15]. In particular, the effects of several routine processing steps on line-shapes, including window functions such as a traditional sine-bell (SINE) and the combined sine-bell - exponential (SEM) function are not readily available. The SEM window function has recently been shown to increase spectral sensitivity and reproducibility in 2D JRES spectra compared with those obtained when processed using SINE [65] and is discussed in section 4.1. In addition, tilting the 2D JRES spectra by 45° and then symmetrising are routine processing steps, not only as part of the calculation of 1D projections of JRES spectra but also for increasing the signal-to-noise ratio in the 2D JRES datasets. Although these processing steps induce a major change to the appearance of the NMR resonances, their effect on the line-shape remains uncharacterized.

This chapter contains a discussion of the line shapes of the 2D JRES spectrum, including the effects of the application of SINE and SEM window functions, comparing simulated and experimental results throughout. The effects of JRES specific processing on the line-shapes of multiplets are also discussed. This may enable any errors of quantitation introduced by the processes of tilting and symmetrising to be calculated. Potential errors of quantification that arise from overlap of the dispersive tails are also investigated. This work has been accepted for publication in the Journal of Magnetic Resonance in Chemistry, forming part of the results in Parsons et al. [48]. This work also provides the basis of the metabolite quantification work discussed in chapter 6.

5.1 Methods and simulation

5.1.1 Experimental NMR spectra

Pure standards of the metabolites adenosine 5-monophosphate (AMP), alanine, asparagine, aspartate, fumurate, glutamate, glutamine, glycine, histidine, isoleucine, lactate, leucine and proline (5 mM final concentration; Sigma-Aldrich) were prepared in 100 mM sodium phosphate buffer (in 90% H2O, 10% D2O; pH 7.0) containing 1.0 mM sodium 3-trimethylsilyl-2,2,3,3-d_4-propionate (TMSP) and 0.2% (w/v) sodium azide[1].

All samples were measured on a DRX-500 NMR spectrometer (Bruker BioSpin, Fremont, CA) equipped with a 5 mm cryogenically cooled triple resonance probe, and operated at 500.18 MHz with a sample temperature of 27°C. 1D 1H NMR spectra were obtained using a 30° excitation pulse, 6-kHz spectral width, and 10-s relaxation delay, employing excitation sculpting for water suppression. Sixteen transients were acquired using 32k data points resulting in a 5-min total acquisition time per 1D spectrum. The resulting data were then Fourier transformed, manually phase corrected and calibrated using the TMSP resonance at 0.0 ppm. The spectra were processed using NMRLab[20]

[1]In this chapter, the collection of NMR spectra was performed by S. Tiziani and C. Ludwig. All simulation, processing and analysis was performed by the author

within Matlab (version 7, The Mathworks). 2D ^{1}H JRES NMR spectra were acquired using 16 transients per increment for 32 increments that were collected into 16k data points, using spectral widths of 6 kHz along the direct dimension (i.e., chemical shift axis) and 50 Hz along the indirect dimension (i.e., spin-spin coupling axis). Prior to Fourier transformation in the direct and then indirect dimensions, data were either not apodised, SINE apodised (i.e., multiplied by a sine-bell window function in each dimension), or SEM[28] apodised (i.e., multiplied by a combined sine-exponential function along the direct dimension and by a sine-bell function along the indirect dimension). Next the 2D spectra were tilted by 45° and symmetrised, all using NMRLab. Finally each spectrum was calibrated by setting TMSP to 0 ppm and 0 Hz, on the direct and indirect dimensions, respectively.

5.1.2 Simulated NMR spectra

All simulated spectra were constructed using custom written Matlab code using the line-shapes derived below. Spectral parameters can be found in table A.2 and are consistent for all spectra for ease of comparison between the different processing procedures. The effective spin-spin relaxation value, T_2, was estimated by fitting the known 1D line-shape to the resonance at 3.57 ppm in the 1D ^{1}H NMR spectrum of glycine using a least squares method. The effective relaxation value in the indirect dimension, τ_2 , was estimated in a similar manner, except using the resonance at 3.57 ppm in the 2D ^{1}H JRES NMR spectrum of glycine (while keeping T_2 constant). These resonances were chosen to calibrate the relaxation parameters since glycine is a common metabolite in biological mixtures with a single, well-resolved NMR signal.

5.2 Effect of apodisation on JRES line-shapes

When first acquired, a single, noise-free, free induction decay (FID) in the time domain is described as[15]:

$$s(t_1, t_2) = A \exp(iW_1 t_1) \exp(iW_2 t_2) \exp(-t_2/T_2) \exp(-t_1/\tau_2) \tag{5.1}$$

where t_1 and t_2 are time in the first (indirect) and second (direct) dimensions, A is the amplitude, (W_1,W_2) are the complex frequencies, and τ_2 and T_2 are the effective transverse relaxation times for the first and second dimensions, respectively. Although a typical JRES spectrum of a biological mixture typically exhibits a few hundred signals, only a single resonance is considered at this point for ease of calculation.

Applying a Fourier transform in each dimension then gives the line-shape for a resonance in a JRES spectrum within the frequency domain:

$$\begin{aligned} \mathcal{F}\left(s(t_1, t_2)\right) &= S(w_1, w_2) \\ &= \int_0^{a_1} \exp(-iw_1 t_1) \int_0^{a_2} \exp(-iw_2 t_2)\ s(t_1, t_2)\ \mathrm{d}t_2 \mathrm{d}t_1 \end{aligned}$$

Here, a_1 and a_2 represent the finite acquisition time of the experiment. However, the acquisition time is designed to be sufficiently long that the signal from the experiment is collected. Thus, it can be assumed that all signals decay as $t_1, t_2 \to \infty$. Typically, a JRES experiment is recorded in magnitude mode, where the absolute value of the real and imaginary channels of the spectrometer are combined to produce a spectrum. Less frequently power mode is used (the square of magnitude mode), but all JRES spectra presented in this thesis are magnitude mode. This processing is a necessary step as analysing a single channel (as is often done in 1D NMR) does not present a usable spectrum, due to the lack of phase information contained in the experiment as a result of the pulse sequence used to separate the spin-spin coupling information from the chemical shift. Hence a simple, non-apodised resonance in a 2D JRES spectrum is represented in the frequency domain

(w_1, w_2) as:

$$s(w_1, w_2) = \frac{T_2\tau_2}{(-1 + iT_2\,(W_1 - w_1))\,(-1 + i\tau_2\,(W_2 - w_2))} \tag{5.2}$$

This is then given in magnitude mode as:

$$|s(w_1, w_2)| = \frac{T_2\tau_2}{\sqrt{(1 + \tau_2^2(W_1 - w_1)^2)\,(1 + T_2^2(W_2 - w_2)^2)}} \tag{5.3}$$

This result is available from many sources[4, 19, 15] and is sometimes written as a product of two Lorentzian line-shapes: the absorptive and dispersive parts (analogous to 1D methods). While experimental methods exist for the acquisition of a JRES spectrum in pure absorption mode[49, 50], these methods suffer from reduced sensitivity and hence are less suitable for metabolomics studies of low concentration metabolite mixtures. Figure 5.1A shows the experimentally acquired glycine resonance at 3.57 ppm, while figure 5.1B shows the corresponding simulated line-shape. These signals show long "tails" that spread far from the peak maxima in both dimensions. These can be a source of interference in areas of high peak density, making spectral deconvolution difficult, as discussed below.

Improved signal-to-noise ratios are often achieved through the use of one of many different *apodisation* (window) functions. Although JRES datasets have traditionally been processed using a sine-bell (SINE) function[71], other functions are also used such as the combined sine-bell and exponential (SEM) function[65]. Since apodisation functions are applied to the FID and subsequently affect the NMR line-shape, it is important to derive the altered line-shape functions for fitting purposes. For SINE apodization, equation 5.1 is multiplied by $\sin(\pi t_i/a_i)$ prior to Fourier transformation of each dimension i, where a_i defines the acquisition time for the experiment:

$$\begin{aligned}
\mathcal{F}\,(s(t_1, t_2)) &= S(w_1, w_2) \\
&= \int_0^\infty \sin\left(\frac{\pi t_1}{a_1}\right) \exp(-iw_1t_1) \int_0^\infty s(t_1, t_2) \sin\left(\frac{\pi t_2}{a_2}\right) \exp(-iw_2t_2)\mathrm{d}t_2\mathrm{d}t_1
\end{aligned}$$

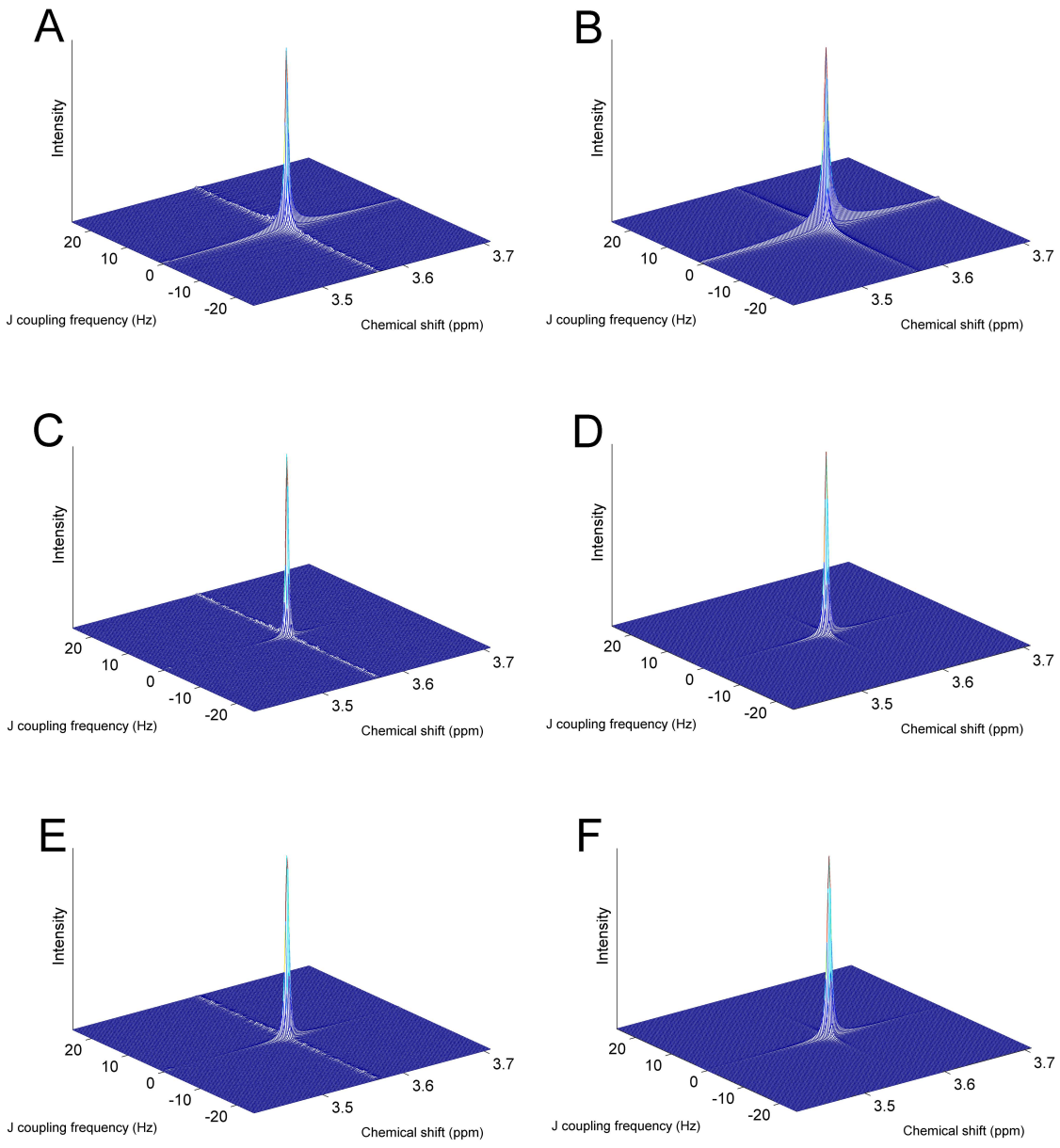


Figure 5.1: Effect of apodisation on the line-shapes of experimental (glycine, 3.57 ppm) and simulated resonances in a 2D JRES spectrum. (A) experimental data, non-apodised; (B) simulated, non-apodised; (C) experimental data, SINE apodised; (D) simulated, SINE apodised; (E) experimental data, SEM apodised, and (F) simulated, SEM apodised.

The modified line-shape in the frequency domain is then:

$$s(w_1, w_2) = \frac{a_1 T_2^2 a_2 \tau_2^2 \pi^2}{(\lambda_1 + i\mu_1(W_1 - w_1))\,(\lambda_2 + i\mu_2(W_2 - w_2))} \tag{5.4}$$

Taking the magnitude then yields the equation:

$$|s(w_1, w_2)| = \frac{a_1 T_2^2 a_2 \tau_2^2 \pi^2}{\sqrt{(\lambda_1^2 + \mu_1^2(W_1 - w_1)^2)(\lambda_2^2 + \mu_2^2(W_2 - w_2)^2)}} \tag{5.5}$$

where $\mu_j = -2a_j^2 T_2$ and $\lambda_j = \pi^2 T_2^2 + a_j^2 - a_j^2 T_2^2 (W_j - w_j)^2$ with $T_2 = \tau_2$ if $j = 1$. Figure 5.1C shows the experimentally acquired and SINE apodised glycine resonance, while figure 5.1D shows the simulated line-shape according to equation 5.5. Here, the peak tails are greatly decreased in both the direct and indirect dimensions. This clearly illustrates the advantage of using an apodisation function, as the truncation of the tails will reduce the likelihood of overlap and interference between closely spaced resonances.

Similarly, the SEM apodisation function can be applied to the FID prior to Fourier transformation. In this case an exponential function is applied along the direct dimension (t_2) and a sine-bell function is applied along the indirect dimension (t_1). The SEM function itself is defined as $\sin(\pi t_2/a_2)\exp(-Lt_2)$ for a given line broadening factor, L, measured in Hertz. The line broadening factor is known to effect the width of the resultant spectral peaks.

$$\begin{aligned} \mathcal{F}\left(s(t_1, t_2)\right) &= S(w_1, w_2) \\ &= \int_0^{a_1} \sin\left(\frac{\pi t_1}{a_1}\right) e^{-L \cdot t_1} e^{-iw_1 t_1} \int_0^{a_2} s(t_1, t_2) \sin\left(\frac{\pi t_2}{a_2}\right) e^{-iw_2 t_2} \mathrm{d}t_2 \mathrm{d}t_1 \end{aligned}$$

The SEM line-shape in the frequency domain is then:

$$s(w_1, w_2) = \frac{a_1 T_2^2 a_2 \tau_2^2 \pi^2}{(\lambda_1 + i\mu_1(W_1 - w_1))\,(\lambda_2 + i\mu_2(W_2 - w_2))} \tag{5.6}$$

Taking the magnitude then yields the equation:

$$|s(w_1, w_2)| = \frac{a_1 T_2^2 a_2 \tau_2^2 \pi^2}{\sqrt{(\lambda_1^2 + \mu_1^2(W_1 - w_1)^2)(\lambda_2^2 + \mu_2^2(W_2 - w_2)^2)}} \tag{5.7}$$

Clearly resultant line-shape is the same as the SINE case (equation 5.5), but here, the effect of the SEM function changes the parameters in the direct dimension to $\mu_2 = -2a_2^2 T_2(1 + T_2 L)$ and $\lambda_2 = \pi^2 T_2^2 + a_2^2 - a_2^2 T_2^2 (W_2 - w_2)^2 + 2a_2^2 L T_2 + a_2^2 L^2 T_2^2$ whilst μ_1 and λ_1 remain unchanged. Notice that if $L = 0$, the SEM line-shape reduces to the SINE case and so the SINE apodisation can be considered a special case of the SEM window function. The SEM apodised signal (see figure 5.1E for experimental data and figure 5.1F for simulated line-shape) appears similar to the SINE apodised signal, again reducing the possible interference between the tails of closely spaced resonances when compared to the non-apodised line-shape.

Overall, figure 5.1 compares each simulated line-shape with a similarly processed experimentally acquired NMR resonance. All signals have been normalised to a peak height of one to facilitate comparison of the peak shapes. Qualitatively, there is good agreement between the line-shapes of each of the simulated and experimental signals. The total areas of the simulated and experimental signals were compared in order to assess qualitatively the quality of the fit. For the non-apodised resonance there is an 18.1% difference between the two areas, while the SINE and SEM apodised resonances are each different by only 5.4% and 3.1%, respectively. This error will be over-estimated due to the presence of noise (largest for the non-apodised data) in the experimental spectra, as this is not accounted for in their simulated counterparts. Skyline projections of the SEM apodised experimental data and the simulated line-shape are shown in figure 5.2 (for both direct and indirect dimensions), where it can be seen that there is good agreement in the peak shapes. This confirms that the derived apodised line-shape is an accurate representation of the experimental results. The most significant effect of the apodisation functions is on the length of the peak tails in the direct dimension, with both the SINE and SEM functions greatly

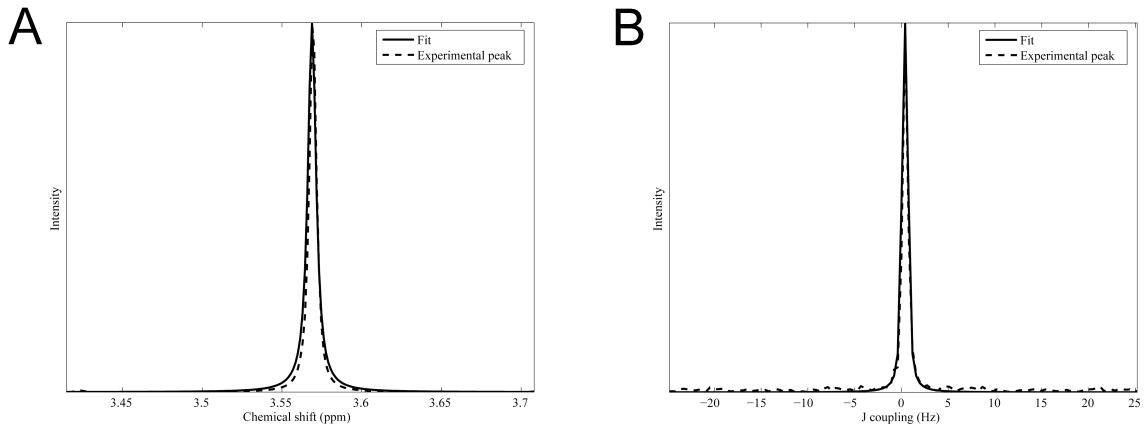


Figure 5.2: Projections of experimental JRES NMR data (dashed line) and simulated line-shapes (solid line) of the SEM apodised glycine resonance at 3.57 ppm. (A) Skyline projection onto the spin-spin coupling axis, and (B) skyline projection onto the chemical shift axis.

increasing the attenuation rate of the signal. This in turn will reduce peak overlap in highly congested spectra as is typical for the analysis of complex biological samples. As noted above, the SEM and SINE apodisation functions are similar, but produce different profiles along the direct dimension dependent upon the line broadening factor. For the line broadening factor used here ($L = 0.5$), the peak tails are more prominent for the SINE apodised resonance when compared with the SEM apodised resonance.

5.3 Effects of JRES specific processing

Figure 5.3 illustrates the effects that tilting and symmetrising (see section 1.3.5) have upon the simulated line-shape of a single SEM-apodised resonance. The twisted line-shape that results from the spectral tilting (figure 5.3B) appears similar to the untilted signal (figure 5.3A). However, the symmetrised resonance (figure 5.3C) has a wider maximum and the tails in the spin-spin coupling dimension are removed, as the spectrum is now symmetrical about the line of J =0 Hz. Tilting and symmetrising a spectrum are particularly important processing steps prior to projecting the spectrum onto the direct dimension to form a 1D pJRES dataset. This type of 1D representation of the spectrum has proven particularly useful in a number of metabolomics analyses[71, 75] due to the

significant reduction of spectral congestion as compared with the 1D spectrum of the same sample. This simplification of spectra is due to each multiplet in the 2D spectrum being collapsed into a single resonance in the 1D projection due to the tilting process. However, as is evident from figure 1.8B, the tails of the resonances are not aligned after tilting. The process of symmetrization then removes the non-aligned tails, truncating the signals and broadening the maxima. This has the potentially problematic consequence of decreasing the intensity of a resonance, as peaks lying upon the line $J = 0$ are often truncated as a result of the averaging of the data points either side of the maxima which can been seen in figure 5.3C). This could in turn adversely affect our ability to determine the concentration of a particular metabolite in a mixture. If symmetrising uniformly affects the resonances across the entire spectrum then the relative areas remain constant. If, however, the different peak structures are not uniformly affected it would become much harder to calculate the relative areas of the metabolite signals and could potentially induce quantification errors. Therefore a quantitative investigation of the effects of tilting and symmetrising on the areas of a singlet, doublet and triplet NMR resonances is necessary.

Table 5.1 lists the total areas of these three types of resonances after each processing step; the total areas of the apodised singlet, doublet and triplet have each been set to unity to facilitate comparison. Clearly the signal areas are unaffected by tilting, but the area decreases to less than half the original value upon symmetrising the spectrum. Importantly, the intensities of the singlet, doublet and triplet signals are affected in an almost identical manner, differing by less than 2% in total area. This reduction in intensity is not obstructive to quantifying metabolites in a sample, since the area relative to the internal standard (e.g. TMSP) is the most important factor in determining concentration. Furthermore this error will have minimal impact on the (semi)quantification conducted in NMR metabolomics studies, where typical (median) peak relative standard deviations of 12.5% (from technical variation) and 20-30% (from inter-individual biological variation) have been reported[46]. See chapter 2.2 for examples.

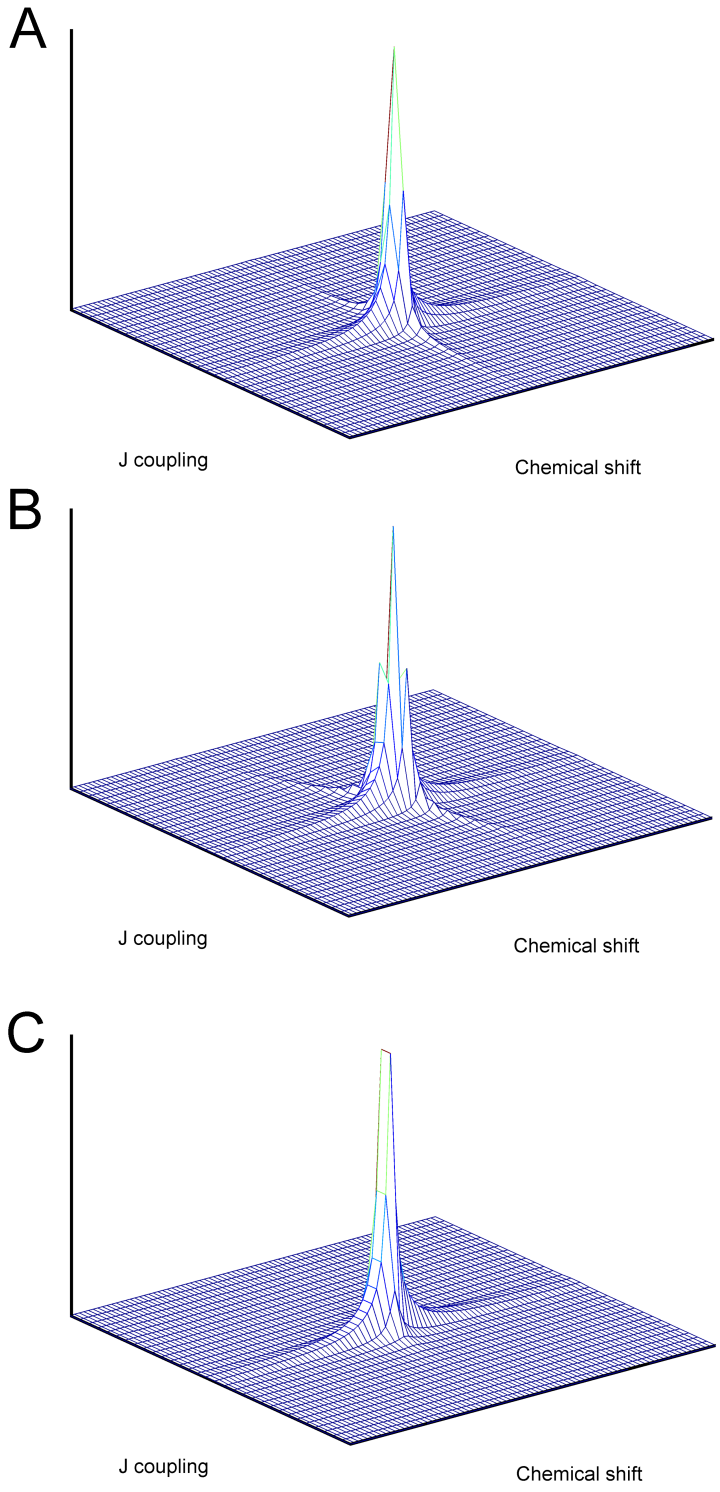


Figure 5.3: Effects of JRES specific spectral processing on a simulated SEM apodised NMR signal. Full three-dimensional line-shape showing: (A) SEM apodised resonance in magnitude mode; (B) application of tilting function to signal in A; and (C) application of symmetrisation function to signal in B.

Stage of processing	Singlet	Doublet	Triplet
SEM apodised	1	1	1
SEM apodised + tilting	1	1	1
SEM apodised + tilting + symmetrisation	0.462	0.476	0.481

Table 5.1: Total spectral areas of different NMR resonances before and after JRES specific processing. For comparison, the areas of the SEM apodised (but otherwise unprocessed) singlet, doublet and triplet resonances are set to unity.

Another noteworthy point concerns the very small features that appear at the foot of closely spaced signals that are introduced by symmetrisation. As shown in figure 5.4, the baseline does not return to zero between the resonances in either the simulated (figure 5.4A) or experimental (figure 5.4B) triplet of isoleucine. These small features arise from the overlapping of the long tails of the three signals in the triplet and are not removed by symmetrisation. This overlapping of peaks is also exhibited by figure 4.3 (page 99). Such features are potentially problematic, and could be classified as signals if a "peak picking" algorithm is applied to the intact 2D JRES spectrum. In terms of projections of JRES spectra, this tail interference does not affect the skyline projection as only the largest resonance (at any given chemical shift value) contributes to the projection. These small signals will, however, affect the intensity of a sum projection by distorting the projected line-shape. This is not a problem for metabolomics studies which to date have utilised skyline (and not sum) projections.

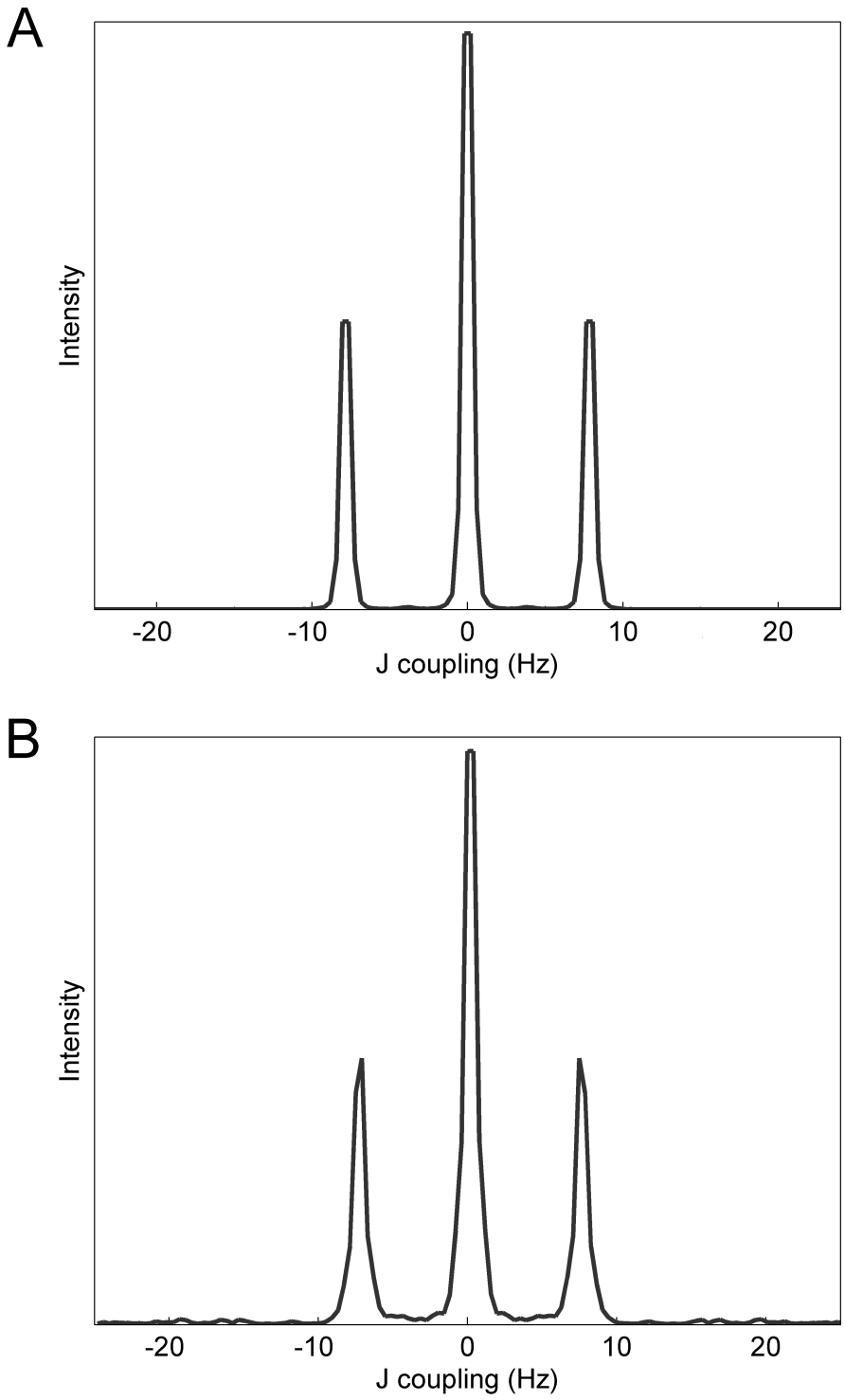


Figure 5.4: Skyline projections of a SEM apodised, tilted and symmetrised triplet onto the J coupling axis (indirect dimension). (A) Simulated line-shape showing slightly raised (non-zero at ±5Hz) baseline between the principal resonances; and (B) experimental data (isoleucine, 0.94ppm) with minor artefacts between the resonances arising from the symmetrisation.

5.4 Conclusions

In this chapter the mathematical functions that describe the line-shape of an NMR resonance within a 2D JRES spectrum under different processing conditions have been determined. This includes for a non-apodised signal as well as following SINE apodisation and the recently highlighted SEM apodisation [65] (see chapter 4). In addition the description of the line-shapes following JRES-specific data processing has also been presented. These results now enable the investigation and development of line-shape fitting approaches for 2D JRES spectra, which is discussed in chapter 6 as part of the examination of the potential quantification of 2D JRES spectra.

The analysis of the specific spectral processing and presentation of JRES spectra has also yielded interesting results. Here, it has been demonstrated that spectral symmetrisation has a major impact upon these line-shapes and reduces the signal intensity approximately by a factor of two. However, this reduction in signal intensity is largely consistent across a singlet, doublet and triplet, and therefore does not appear to be a major source of quantitation error. Symmetrising a spectrum also generates low intensity features near the base of peaks. Although these are very small compared with the main resonance, they could result in a large number of false positives when applying a "peak picking" algorithm.

Chapter 6

Quantitation of JRES spectra

In this chapter, the analytical descriptions of the line-shapes of the 2D JRES spectra established in chapter 5, are used to investigate if JRES spectra are a suitable tool for metabolite quantification. As described previously, quantification of metabolites has many potential uses in metabolomics, and has been attempted with many disparate techniques for NMR spectra[24, 25, 27, 44, 82]. Fitting line-shapes to the spectral peaks is no exception, particularly for 1D spectra [1, 21], where each resonance has a known line-shape. Whilst fitting in the time domain has been attempted [52], this thesis uses only the resonance shapes in the frequency domain, exploiting the unique properties of peak structures found in the JRES experiment.

The investigation described in this chapter initially presents the examination of quantitation errors of closely spaced peaks using synthetically created spectra, to quantify the effect of the long peak 'tails' described previously. Also, using a simple peak picking and fitting routine, a quantification algorithm is tested using a variety of spectra of known mixtures and 'true' biological samples via the use of a library of reference spectra. Finally, the results are compared with those acquired from 1D NMR spectra using commercially available software.

This work forms part of the results accepted for publication in Magnetic Resonance in Chemistry [48].

6.1 Quantitation errors of closely spaced peaks

Resonances do not occur in isolation in 2D JRES spectra; typically a complex biological mixtures may contain many hundreds of signals. Even though these signals are less congested than in 1D NMR experiments, resonances can still overlap with each other in either dimension. The use of magnitude mode in a 2D JRES spectrum exacerbates this problem as the interference becomes non-linear. In particular, the dispersive tails of neighboring resonances can potentially adversely affect the total signal areas, which would introduce quantitation errors. Therefore it is necessary to investigate (and quantify) the effects of overlapping resonances on signal intensity, in particular the dependence on peak spacing. The effects of peak convergence can be readily seen by placing two complex resonances, each of unit area, along the centreline of a spectrum, converting to magnitude mode, and then measuring the total spectral area of the two peaks as the distance between them is varied. Figure 6.1 shows this for each of the apodisation methods described above.

At a large separation distance, d, the total area of the pair of non-overlapping resonances approaches zero error, irrespective of which apodisation function is employed (data not shown). When the spacing is decreased to 25 times the peak full-width-half-maximum (FWHM), each pair of SINE or SEM apodised resonances effectively remain non-interacting, while the long tails of the non-apodised resonances now overlap significantly and induce an ca. 15% error in the total signal area. As d decreases further the quantitative errors increase, with the non-apodised, SINE apodised and SEM apodised resonances showing quite different behaviours. The maximum error for a pair of apodised signals occurs when they are very closely spaced, ca. one FWHM apart, with the greatest error exhibited by the SINE case (ca. 33% error in total signal intensity), and a smaller but sizeable error for SEM apodisation (ca. 20% error). Although the maximum error is

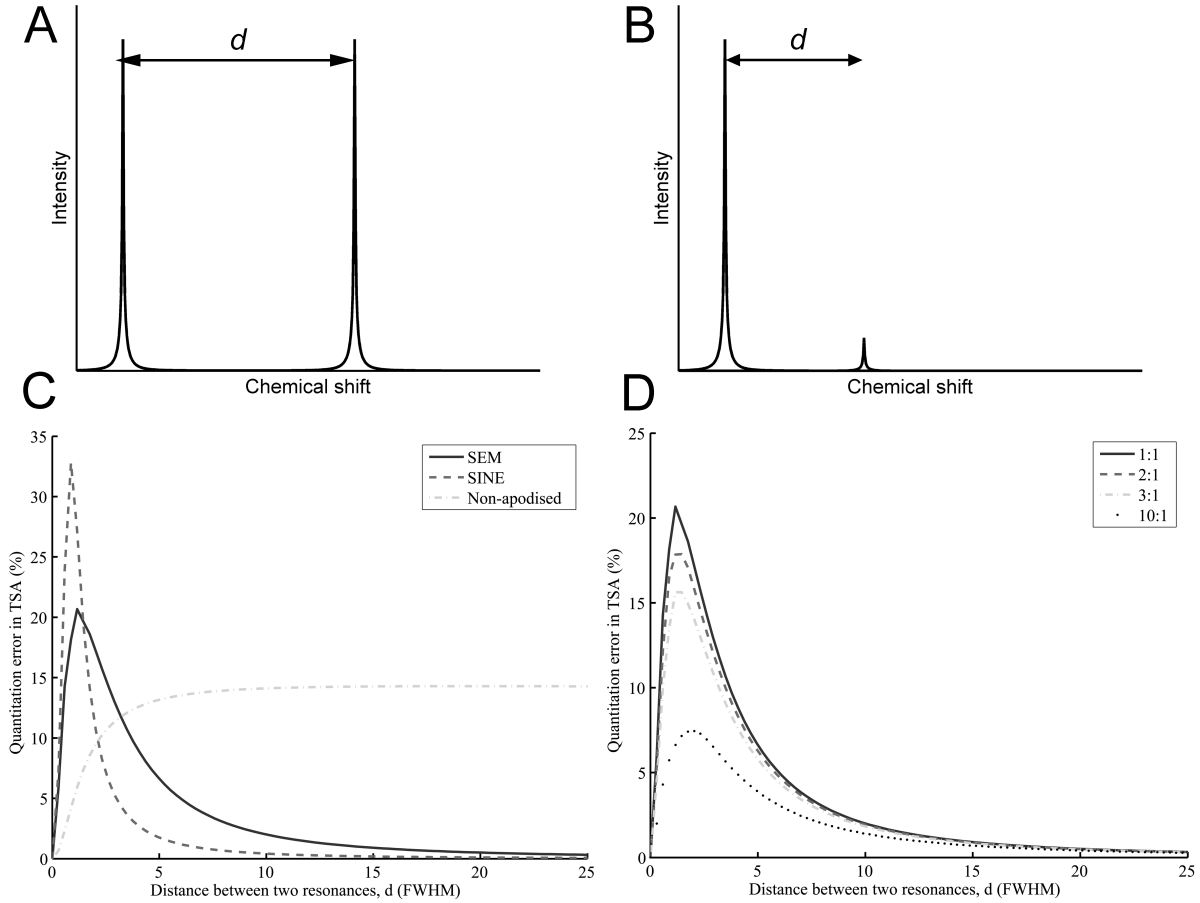


Figure 6.1: Effect of overlapping JRES NMR resonances on total spectral intensity. (A) Schematic representation of two neighbouring, equal (unit) intensity resonances separated by distance d (measured in terms of peak full-width-half-maximum); and (C) total spectral area (TSA) of the two resonances (shown as a percentage of the true value of two) as a function of the distance, d, between the signals. A significant error in signal intensity occurs when the resonances are less than 5 times the FWHM for SINE and SEM apodised signals. (B) Schematic representation of two neighbouring, unequal intensity resonances separated by distance d (measured in terms of peak FWHM); (D) the effects upon the TSA (shown as a percentage of the true value) as a function of the distance, d, between the signals. Ratios of 1:1, 2:1, 3:1 and 10:1 peak intensities are shown. The error in TSA does not decrease linearly with respect to peak height.

less for the SEM case, the range over which this error is $> 5\%$ is greater (than for SINE apodised resonances), spanning 6 times the FWHM.

The possibility of peaks of uneven sizes interacting is also high. To investigate this, spectra were simulated consisting of two SEM apodized peaks of height ratios 1:1, 2:1, 3:1 and 10:1. These results are shown in figure 6.1D. As above, the effect upon the total spectral area of the changes as the peaks move further apart. Although the error in the total spectral area decreases as the intensity of the second peak decreases, it does not do so in a linear manner. For all peak ratios, a distance of over 10 FWHM is required for the error in area to drop below 1%, indicating the need to account for the dispersive tails in all circumstances.

These errors highlight the importance of apodising FIDs measured in JRES experiments, in order to truncate the long dispersive tails. More importantly these results draw attention to the inherent quantitation error associated with the deconvolution of an intact 2D JRES spectrum of a biological sample into its component metabolites. For the case of closely spaced signals (e.g., from one to five FWHM apart), quantitation errors of up to 25% (for SINE) or 17% (for SEM) will occur if the magnitude mode spectrum is deconvoluted using a basis set of processed (i.e. magnitude mode) JRES spectra of individual metabolites. These errors are not sufficiently large to prohibit such an approach since metabolic changes between, for example, control and diseased samples maybe 1.5-fold or greater. Furthermore, a given metabolite may have multiple NMR resonances of which most are well resolved and not affected by the error. However, due consideration of this error will be important in future metabolomics studies that attempt accurate metabolite quantification.

6.2 Line-shape quantification

6.2.1 Experimental methods

Pure metabolite standards of each metabolite (5 mmol/L; Sigma-Aldrich) were prepared as described earlier (section 5.1.1)[1]. Chemically-defined mixtures were created to mimic the NMR-characterised metabolite concentrations of three different sample types: a human cancer cell line [64], an adductor muscle extract from a marine mussel [22] and a fish embryo extract [71]. These mixtures were created using pure metabolite standards of adenosine 5-monophosphate (AMP), alanine, asparagine, aspartate, fumurate, glutamate, glutamine, glycine, histidine, isoleucine, lactate, leucine and proline. Each mixture was created in the same buffer, with a further 1.0 mM TMSP added.

Tissue extracts were derived from 19 male roach (*Rutilus rutilus*) that were exposed to either no (control), low or high doses of a model toxicant (fenitrothion). Briefly, polar metabolites were extracted from testis tissue using a methanol:chloroform:water method and Precellys 24 homogeniser (Stretton Scientific, Stretton, UK) as described previously [81]. Immediately prior to NMR analysis, the dried polar extracts were re-suspended in sodium phosphate buffer (as above) containing 0.5 mM TMSP.

The experimental procedure to acquire the 2D JRES spectra is described previously in section 5.1.1. For the 1D experiments, all spectra were obtained using a 30° excitation pulse, 6-kHz spectral width, and 10-s relaxation delay, employing excitation sculpting for water suppression [68]. Sixteen transients were acquired using 32k data points resulting in a 5-min total acquisition time per 1D spectrum. The resulting data were then Fourier transformed, manually phase corrected and calibrated using the TMSP resonance at 0.0 ppm. These spectra were then processed using TopSpin (v1.3; Bruker). Finally, Selected metabolites in the 1D ^{1}H spectra were quantified using Chenomx NMR Suite (version 5.1;

[1]All experimental spectra were recored by Stefano Tiziani and Adam Hines. The author subsequently processed analysed the data

Chenomx Inc., Edmonton, Canada).

6.2.2 Method of area comparison

All line-shape based quantification was carried out using custom written Matlab code, using tilted but not symmetrised spectral data to minimise all known sources of error (see section 5.3). For each metabolite to be quantified, the line-shape method requires a 2D JRES spectrum of the pure metabolite (at a known concentration; referred to as the "library" spectrum) as well as the spectrum of the sample requiring quantification (the "sample" spectrum). It is important that these JRES NMR spectra are acquired and processed in an identical manner. In particular we have found that a minimum of 32 increments (in the indirect spin-spin coupling dimension) are required during acquisition of the free induction decay (FID) to ensure that the spectral resolution is sufficient for the analysis (see section 4.2). For the library and sample spectra, the concentrations of the internal standard (e.g. TMSP) must also be known, but may be unequal.

After routine processing to yield intact 2D JRES spectra, the resonance arising from the internal standard (see section 1.3) was located, the line-shape was fit and the area of the resonance calculated, for both the sample and library spectra.

The fits in this section are estimated by calculating the maximum intensity of the structure of interest in the magnitude mode spectrum, then calculating the line-shape of a theoretical complex resonance of that intensity. By combining the real and imaginary parts of this theoretical resonance a simulated, noiseless magnitude mode structure can be calculated. This ensures that when simulating peak structures which consisted of multiple resonances, dispersive tail effects are included (described in section 6.1). By simulating the resonance structures of both the library and sample spectra separately, the effects of noise are removed and the structures are subject to a fair comparison.

Next, a resonance structure was selected for the metabolite to be quantified, based upon the following criteria in both the sample and library spectra: the resonance structure is at least 10 times the noise threshold; it is not subject to large pH-induced changes in chemical shift; and ideally it is present in a region of low spectral congestion in both spectra. After the resonance structure is chosen, a simulated synthetic spectrum of both the library and sample spectra are created using the same method as described above, placing as many line-shapes as necessary to create the entire structure (i.e. singlet, doublet, etc) before calculating the final magnitude mode spectrum.

Once the four magnitude mode synthetic spectra have been created, the total spectral area of each is calculated and finally the metabolite is quantified using the following equation:

$$x = \frac{yM_L}{n_L T_L} \cdot \frac{n_s T_s}{M_S} \tag{6.1}$$

where x is the quantity of the unknown metabolite concentration in the sample, y is the known concentration of the pure metabolite in the library spectrum, T_L and T_S are the line-shape derived areas of the internal standard in the library and sample spectra respectively, and n_L and n_S are the known concentrations of internal standard in the library and sample spectra. Finally, M_L and M_S are the line-shape derived areas of the metabolite structure in the library and sample spectra.

6.2.3 Chemically defined samples

Three synthetic samples were created to mimic three disparate sample types: (human) cell extract; marine (mussel) tissue; and fish (medaka) embryo, which are described in more detail in section 6.2.1. To evaluate the effectiveness of 2D JRES compared with standard methods, 1D spectra were quantified using NMR Suite (Chenomx Inc) to provide a benchmark. NMR Suite is a commercially available semi-automated program that uses a fitting method to estimate metabolite quantities. The program displays the spectrum

of the metabolite, where users can superimpose the simulated profile of any of its library metabolites in real-time. Users can manipulate the chemical shifts of each resonance structure of a metabolite independently up to pre-defined limits to account for pH induced shifting. Altering the intensity of one resonance structure alters the intensity of each peak in that metabolite, changing the estimated quantity of the metabolite in the sample. Adding multiple metabolites yields an overall fit, generated in a linear manner, helping the user to estimate the intensity peaks in of highly congested areas. The results can be found in tables 6.1, 6.2 and 6.3.

		1D analysis		JRES analysis	
Metabolite	Gravimetric quantity (μM)	Measured quantity (μM)	Absolute error	Measured quantity (μM)	Absolute error
AMP	73.70	60	18.59%	94.27	27.91%
Alanine	42.20	54.3	28.67%	35.53	15.81%
Asparagine	99.10	64	35.42%	84.43	14.80%
Aspartate	28.90	29	0.35%	30.49	5.51%
Fumarate	2.50	0.8	68.00%	1.14	54.36%
Glutamate	570.50	545.8	4.33%	513.59	9.98%
Glutamine	130.60	27.3	79.10%	288.33	120.77%
Glycine	261.60	146.5	44.00%	148.62	43.19%
Histidine	20.50	4.3	79.02%	35.74	74.33%
Isoleucine	42.10	25.9	38.48%	34.63	17.75%
Lactate	355.60	254.4	28.46%	347.54	2.27%
Leucine	40.60	22.9	43.60%	28.39	30.08%
Proline	180.50	145.2	19.56%	126.99	29.64%
		Mean error	40.49%		38.79%
		Standard deviation	28.40%		36.70%

Table 6.1: The quantification results of the chemically defined cell extract sample

		1D analysis		JRES analysis	
Metabolite	Gravimetric quantity (μM)	Measured quantity (μM)	Absolute error	Measured quantity (μM)	Absolute error
AMP	387.10	202.2	47.77%	315.63	18.46%
Alanine	2524.60	3945	56.26%	2609.84	3.38%
Asparagine	471.70	357.6	24.19%	380.90	19.25%
Aspartate	464.00	460	0.86%	566.26	22.04%
Fumarate	12.60	7.8	38.10%	8.65	31.34%
Glutamate	875.90	763	12.89%	1130.15	29.03%
Glutamine	688.50	260.2	62.21%	3842.82	458.14%
Glycine	12547.90	8294.7	33.90%	8398.95	33.06%
Histidine	176.70	38.8	78.04%	169.51	4.07%
Isoleucine	58.10	39.4	32.19%	61.63	6.08%
Lactate	310.00	214.1	30.94%	432.02	39.36%
Leucine	85.80	52.9	38.34%	60.76	29.18%
Proline	215.30	188.3	12.54%	122.75	42.99%
		Mean error	34.00%		69.53%
		Standard deviation	23.07%		137.14%

Table 6.2: The quantification results of the chemically defined mussel muscle sample

		1D analysis		JRES analysis	
Metabolite	Gravimetric quantity (μM)	Measured quantity (μM)	Absolute error	Measured quantity (μM)	Absolute error
AMP	324.25	167.20	48.43%	351.86	8.52%
Alanine	938.30	978.20	4.25%	676.58	27.89%
Aspartate	448.40	260.70	41.86%	289.35	35.47%
Glutamate	2457.30	1712.40	30.31%	1673.07	31.91%
Glutamine	259.60	64.30	75.23%	975.50	275.77%
Histidine	162.30	38.00	76.59%	172.12	6.05%
Isoleucine	106.90	55.30	48.27%	72.43	32.25%
Lactate	420.60	264.80	37.04%	402.85	4.22%
Leucine	327.30	158.60	51.54%	219.97	32.79%
Proline	371.80	210.90	43.28%	195.89	47.31%
		Mean error	50.52%		58.22%
		Standard deviation	16.96%		89.13%

Table 6.3: The quantification results of the chemically defined medaka embryo sample

	1D analysis		JRES analysis	
	Line of best fit	p value	Line of best fit	p value
Cell extract	$0.88x - 19.04$	< 0.001	$0.87x + 12.72$	< 0.001
Medaka embryo	$0.78x - 47.35$	< 0.001	$0.61x + 147.81$	0.0012
Mussel muscle	$0.68x + 158.43$	< 0.001	$0.65x + 450.46$	< 0.001

Table 6.4: Table of correlation p values of estimated and known values and equations of best fit lines of the three chemically defined samples depicted in figure 6.2

The results are also shown pictorially in figure 6.2, alongside the benchmark 1D quantitation, where it is clear that both the 1D and 2D JRES results are strongly correlated with the gravimetrical quantity and yield comparable average results. It can also be seen that the 1D analysis underestimates the metabolites whilst the 2D JRES analysis overestimates the known quantity. Table 6.4 details the lines of fit and p values of each of the samples.

The results of the JRES quantification exhibit a large range of accuracies, ranging from approximately 2% to 458% error. The reason for this range in error of the quantitation is not immediately obvious, because, as shown in figure 6.3, there is no correlation between the metabolite quantity and the error. The results of a correlation test confirm this, with $r_{fish} = 0.3486$, $r_{mussel} = 0.1835$ and $r_{cell} = -0.1390$, all reporting no correlation between the two variables. Nor is the error metabolite dependent; for example, the error in quantifying alanine is 15.81% for the cell extract, 3.38% for the mussel spectrum and 27.89% for the medaka embryo.

However, it is clear that certain metabolites exhibit large amount of error across each of the three samples. For example, glutamine exhibits the highest absolute error of any metabolite in the synthetic mussel sample (458.14%), but the error also remains high for the synthetic cell extract (120.77%) and synthetic medaka embryo (275.77%) spectra. On examination of the peaks used to quantify the metabolite (figure 6.4) in the synthetic medaka embryo-like sample, the spectrum segment is so congested that it is difficult to see the triplet structure of glutamine (at 3.785ppm) between the two quartet structures

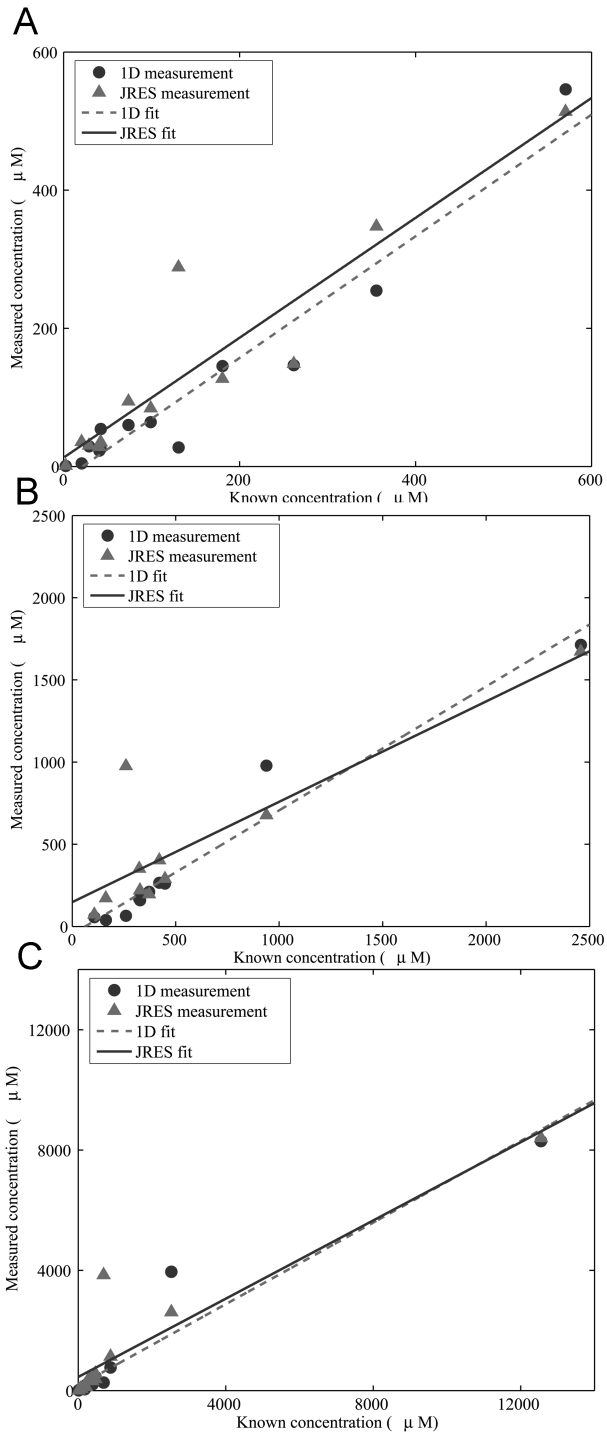


Figure 6.2: Graphical results of the quantification of the three chemically defined samples. Here, each point represents the estimated (via the line-shape method) versus known quantity of each of the metabolites in the samples. Part (A) shows the results of the cell line data, (B) the medaka embryo data and (C) the mussel muscle data. For each sample, the results of the 2D JRES analysis (light grey triangles, dashed line) and the 1D analysis (dark grey circles, solid line) are shown, along with their lines of best fit.

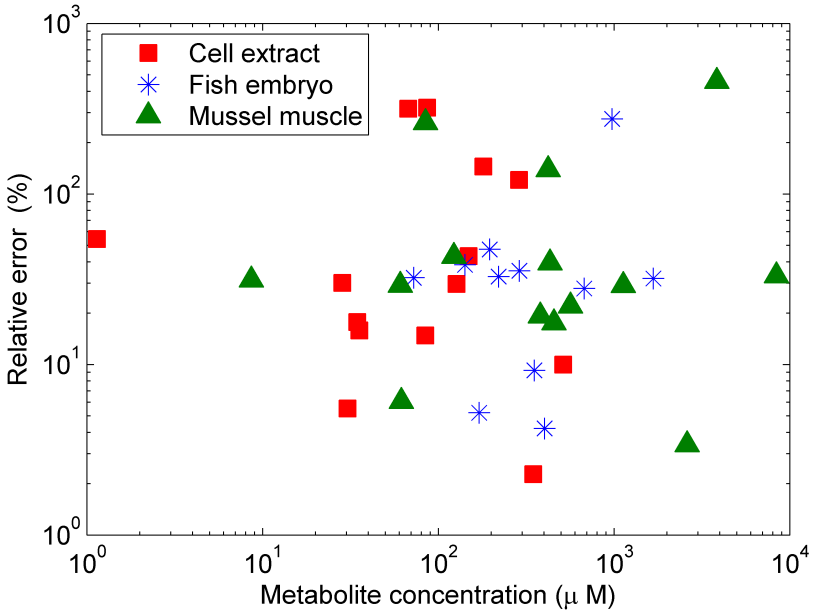


Figure 6.3: Relative error versus metabolite concentration (shown on a $\log_{10}$ scale for clarity), for each of the three chemically defined ('synthetic') samples. Clearly, there is no correlation between the two variables.

on either side. Investigating other metabolites exhibiting high error in both the synthetic mussel muscle sample (phenylalanine, figure 6.5) and the synthetic cell extract sample (histidine, figure 6.6) both reveal congested sections of spectra. These results suggest that spectral congestion and potential mis-assignments are the largest sources of error. They also suggest that the intensity of the peaks is not correlated to the degree of error.

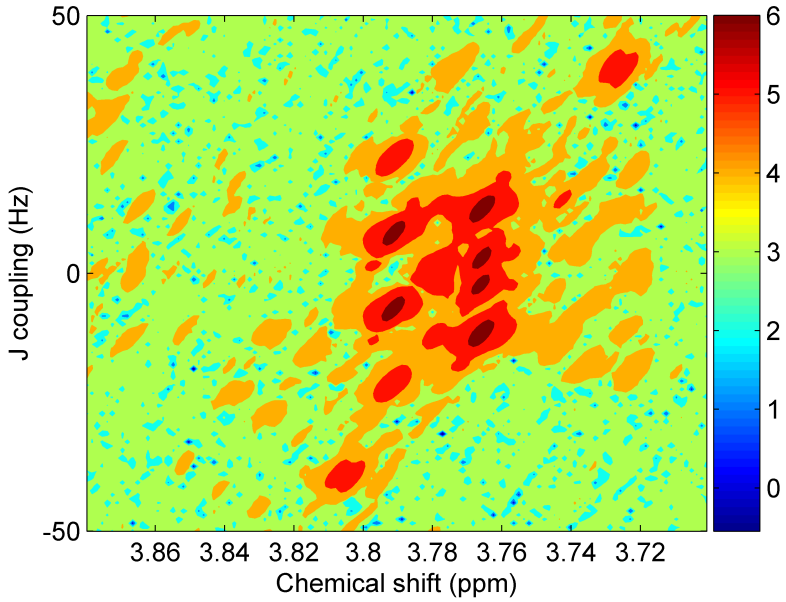


Figure 6.4: Log_{10} plot of the chemically defined fish embryo spectrum at 3.7 - 3.88 ppm. The glutamine structure of interest is a triplet centred at 3.785 ppm, but this is obscured by the more intense structures on either side.

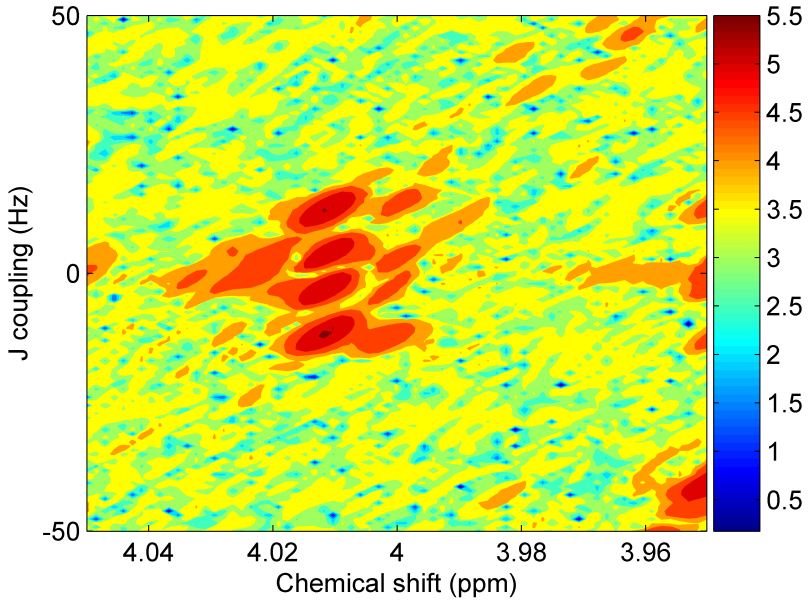


Figure 6.5: Log_{10} plot of the chemically defined mussel spectrum at 3.95 - 4.05 ppm. The phenylalanine structure of interest is a quartet at 4.0 ppm, and is overshadowed by a larger quartet at 4.01 ppm.

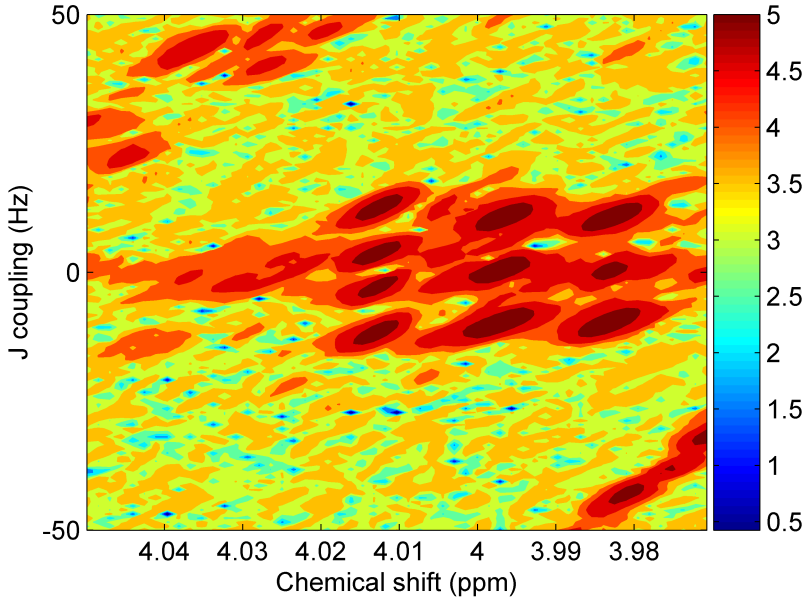


Figure 6.6: Log_{10} plot of the chemically defined cell extract spectrum at 3.97 - 4.05 ppm. The histidine structure of interest is a quartet at 4.0 ppm, but is not present here in this case. It is possible that the structure has shifted to 4.01 ppm, but also, due to the width of the peak in the triplet at 4.0 ppm, it is also possible that other resonances have obscured the middle peaks, presenting a triple-like appearance.

6.2.4 Biological samples

In order to further validate the line-shape fitting approach with NMR spectra of actual biological samples, as well as to extend investigations of the effects of closely spaced peaks, resonances arising from three metabolites that experienced differing degrees of congestion were selected: the well resolved lactate doublet (at 1.33 ppm, figure 6.7), the mildly congested valine doublet (at 3.78 ppm, figure 6.8), and the heavily congested phosphocreatine singlet (at 3.96 ppm, figure 6.9), across 19 spectra acquired from male roach tissue samples subjected to differing concentrations of a model toxicant (see section 6.2.1). The full results of this analysis are presented in tables 6.5 (lactate), 6.6 (valine) and 6.7 (phosphocreatine).

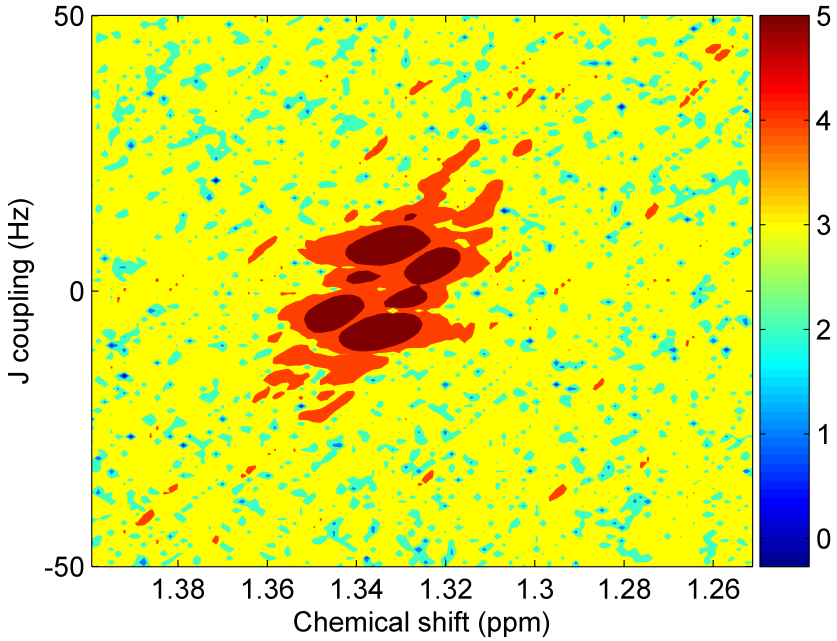


Figure 6.7: Log_{10} plot of the spectrum section showing the lactate resonances used to estimate the concentration (control sample spectrum 1). Here it can be seen that the resonances do not suffer from interference from other resonances present. Note that this spectrum has been tilted, but not symmetrization, which results in extra peaks (usually removed during symmetrization) in the familiar doublet.

The correlation between the traditional 1D measurements (using Chenomx NMR Suite) and those from the 2D JRES line-shape fitting of the JRES spectra is strong for all three

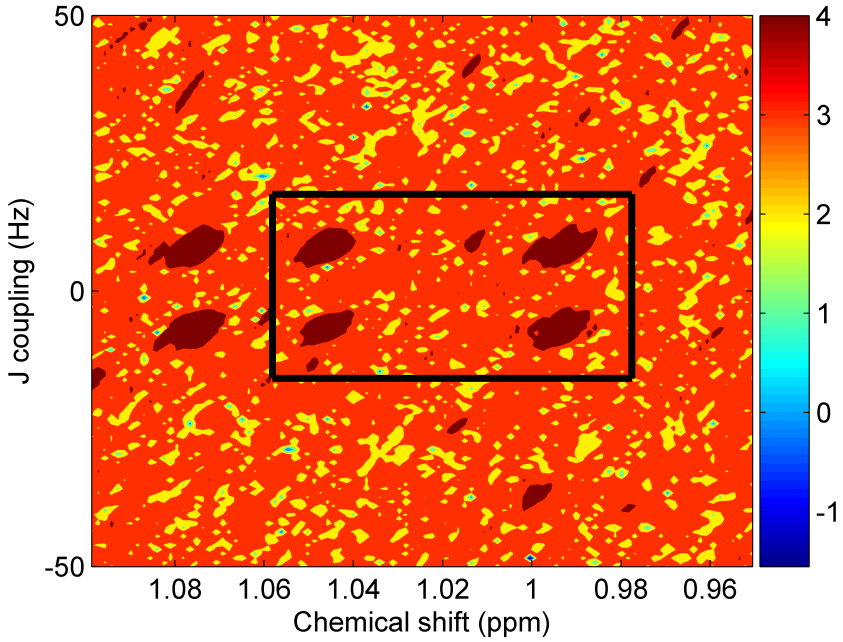


Figure 6.8: Log_{10} plot of the spectrum section showing the valine resonances used to estimate the concentration are enclosed in the black rectangle (control sample spectrum 1). Here it can be seen that other resonances are close by (at 1.08ppm) and hence the valine structure suffers from some interference.

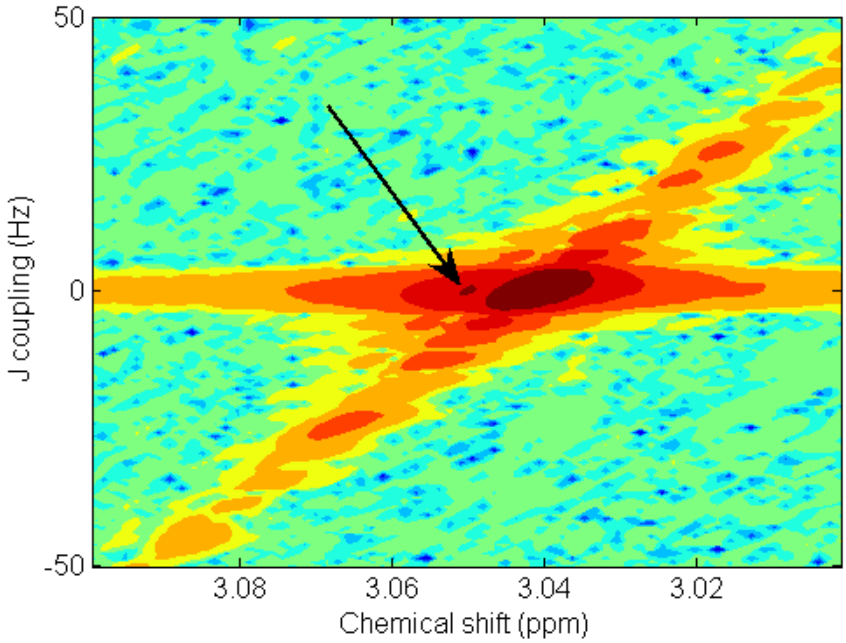


Figure 6.9: Log_{10} plot of the spectrum section showing the phosphocreatine resonance (black arrow) used to estimate the concentration (control sample spectrum 1). Clearly, the tiny phosphocreatine resonance is overshadowed by a much larger (creatine) resonance making peak identification and deconvolution very difficult.

Sample	Class	1D est. (μM)	LS* est. (μM)	Error (%)
1	control	152.4	145.50	1.90
2	control	36.4	31.08	14.61
3	control	75.2	85.52	13.73
4	control	101.2	111.30	9.98
5	control	37.1	168.86	355.16
6	control	45.4	56.47	24.39
7	low dose	51.4	54.75	6.51
8	low dose	35.7	38.75	8.55
9	low dose	29.3	32.83	12.04
10	low dose	36.9	28.67	22.29
11	low dose	88.2	112.51	27.56
12	low dose	54.7	71.19	30.15
13	low dose	76.2	94.95	24.60
14	high dose	66.1	80.40	21.64
15	high dose	59.5	73.00	22.69
16	high dose	47.5	47.11	0.81
17	high dose	80.2	85.80	6.98
18	high dose	50.6	50.21	0.77
19	high dose	67.3	78.03	15.95

Table 6.5: Quantitation results of lactate in the roach tissue samples. *Estimated quantification via line-shape method.

Sample	Class	1D est. (μM)	LS* est. (μM)	Error (%)
1	control	18.7	19.62	4.94
2	control	6.0	5.31	11.51
3	control	7.6	7.88	3.70
4	control	18.6	21.98	18.20
5	control	4.5	22.72	404.82
6	control	6.8	6.60	2.95
7	low dose	3.8	4.62	21.56
8	low dose	5.4	6.32	17.10
9	low dose	3.6	4.42	22.89
10	low dose	1.7	2.28	33.92
11	low dose	8.5	17.03	100.33
12	low dose	5.1	8.04	57.57
13	low dose	9.7	10.36	6.82
14	high dose	6.7	9.53	42.28
15	high dose	5.7	7.24	27.00
16	high dose	6.0	7.84	30.68
17	high dose	8.4	10.79	28.42
18	high dose	4.8	6.28	30.92
19	high dose	5.9	6.33	7.26

Table 6.6: Quantitation results of valine in the roach tissue samples. *Estimated quantification via line-shape method.

Sample	Class	1D est.(μM)	LS* est. (μM)	Error (%)
1	control	166.8	232.50	39.39
2	control	34.9	40.21	15.21
3	control	152.0	268.87	76.89
4	control	17.2	136.48	679.86
5	control	100.0	261.29	161.29
6	control	22.0	94.84	331.09
7	low dose	282.4	203.41	27.97
8	low dose	24.6	23.00	6.51
9	low dose	25.8	29.45	14.15
10	low dose	16.0	99.15	519.70
11	low dose	213.0	187.55	11.95
12	low dose	200.7	226.49	12.85
13	low dose	419.0	197.08	52.96
14	high dose	53.3	200.07	275.37
15	high dose	171.1	190.13	11.12
16	high dose	152.2	348.63	129.06
17	high dose	169.9	202.28	19.06
18	high dose	361.6	209.04	42.19
19	high dose	330.0	308.19	6.61

Table 6.7: Quantitation results of phosphocreatine in the roach tissue samples. *Estimated quantification via line-shape method.

metabolites, shown in figure 6.10. Table 6.8 details the lines of fit and p values from the correlation analyses. From figure 6.10, it can be seen that the line of best fit accurately describes the spread of the data points, indicating a linear correlation between the 2D JRES and 1D quantitation methods. However, it can also be seen that none of the lines of best fit indicates a one-to-one correspondence (i.e. not the line $y = x$). Each of lines has a positive gradient larger than unity, from which may be inferred that the 2D JRES spectra produce a consistently larger estimate than the 1D spectra. The non-zero intercept also confirms that the two experiment types produce differing estimations of the metabolite quantity. This finding is consistent with the results from the chemically defined samples, where the 2D JRES analyses also predominately overestimated the concentrations compared with the 1D approach. Note that each of the values produced is both estimates of the metabolite quantity, unlike for the chemically defined samples discussed in section 6.2.3.

	Mean absolute deviation (%)	Line of best fit	Correlation p value	Degree of congestion
Phosphocreatine	61.24 ± 26.63	$1.92x - 32.29$	< 0.001	Overlapping peak tails
Valine	27.03 ± 23.38	$1.19x + 0.48$	< 0.001	3.5 FWHM to nearest peak
Lactate	15.47 ± 9.71	$1.15x - 1.33$	< 0.001	17 FWHM to nearest peak

Table 6.8: Correlation between the measured metabolite concentrations from analysis of 1D (using Chenomx) and 2D JRES NMR spectra (using line-shape fitting), for the 19 roach gonad samples. The mean absolute deviation between the 1D and 2D JRES measurements, line of best fit, p value and approximate degree of congestion in the region of the analysed signal are presented for each metabolite.

As expected, the most highly congested peak, arising from phosphocreatine, exhibited the largest mean absolute deviation between the JRES and 1D measurements of 61.24%. Again, this value does not represents only an approximate 'error', since the actual metabolite concentrations in these samples are unknown. Quantification of the valine resonances produced a mean absolute deviation between the 2D JRES and 1D spectra of 27.04%, while for the most highly resolved resonances, from lactate, the mean absolute deviation was only 15.47%. Clearly the degree of spectral congestion impacts in a significantly different manner upon the quantification of resonances in 1D to 2D JRES spectra.

Overall these results demonstrate that the JRES line-shape fitting approach can be applied to NMR spectra of biological samples, yielding comparable results to traditional 1D spectral integration.

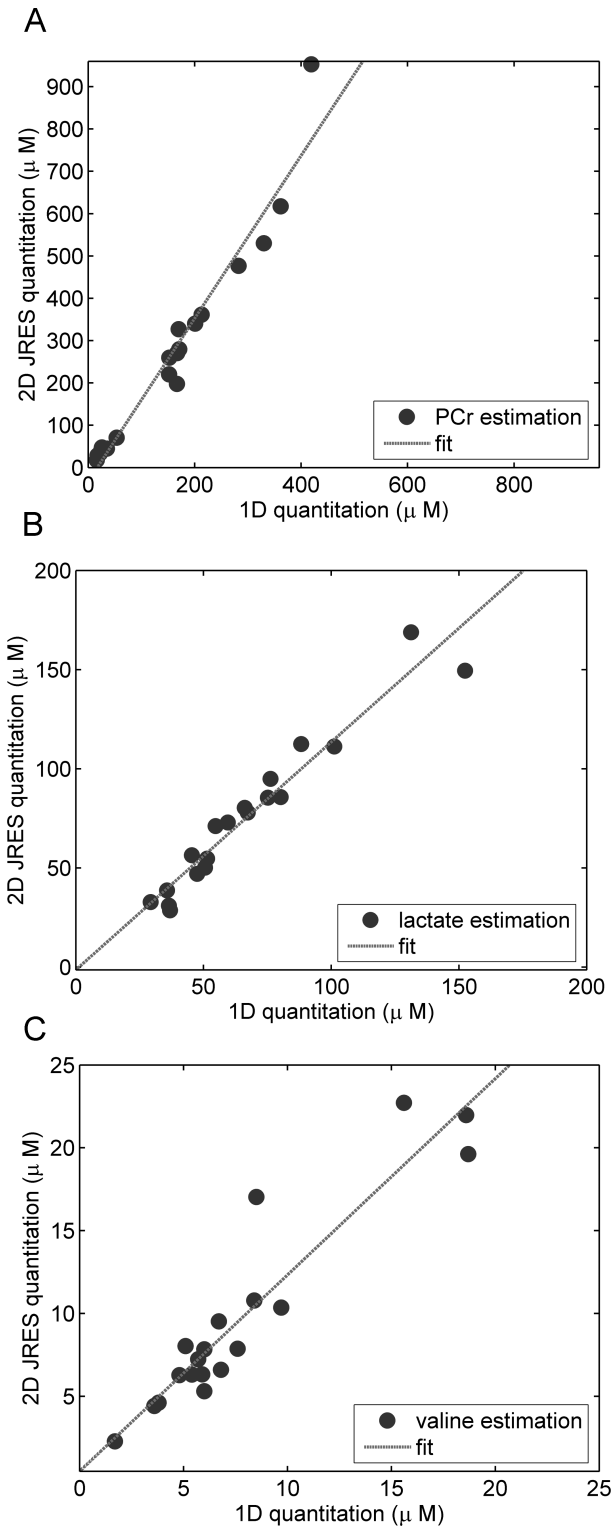


Figure 6.10: Graphical results of the quantification of the roach samples comparing 1D and 2D JRES quantification. (A) Phosphocreatine (PCr); (B) lactate and (C) valine results. For each metabolite, 19 measurements have been made and a line of best fit added.

6.3 Summary and further work

In this chapter, the quantification of metabolites in JRES spectra via line-shapes has been examined for both applicability and effectiveness. By far the most significant error discovered arises from the overlapping of dispersive tails in the 2D spectra, which highlights the importance of apodising the JRES dataset. In order to minimise this error for very closely spaced resonances, it is recommend that FIDs are SEM apodised, which is consistent with the earlier results of section 5.2 which showed that the SEM function is the preferred apodisation method when considering signal-to-noise levels [65].

Using both chemically defined and biologically acquired spectra the effectiveness of the approach has produced mixed results. Whilst the accuracy and reproducibility of the quantitation of lactate in the roach samples, for example, have confirmed the potential of the method; the failure to deconvolve correctly the phosphocreatine peak and confidently identify histidine illustrate that there are many issues that must be still addressed for both 2D JRES quantitation and NMR in general. Primarily, the issue of peak identification must be resolved before this method can be subject to widespread usage. Due to peak shifting and spectral contamination, correctly assigning metabolite to specific resonances becomes a difficult, non-trivial issue. However, there are many different avenues which this thesis has not investigated. For example, 2D JRES spectra can be effectively treated as an image or 3D surface - many algorithms are specifically designed to extract interesting features from these data formats and are extensively studied and (see Baldock and Graham [3], for example).

The estimation of peak area is also a non-trivial issue. A problem specific to 2D JRES spectra is that the peak maxima are intrinsically convolved with other peaks. This is since the resonance tails are so widely spaced, as described in section 6.1, hence may lead incorrect estimation of peak area. Whilst it would be possible to fit all peaks in the complex domain (where peaks are linearly constructed), this approach would yield an ill-

defined system since the complex spectrum cannot be experimentally defined. However, it is unclear if this would provide sufficient error to change significantly peak quantification, or if the noise present in all NMR spectra would completely obscure these errors.

Other avenues of further work include investigation into the effects of the choice of leaving the spin-spin relaxation parameter constant throughout the analysis. This choice was initially made to simplify the analysis as the spin-spin relaxation remains relatively constant for many metabolites, but accounts for an unknown amount of quantitation error. Finally, whilst the issue of resolution in the indirect dimension has been discussed during this work in section 4.2, the use of very high resolution in the indirect dimension (i.e. over 32 increments) has not been examined. Nor has the use of increased resolution as an aid to peak picking. By increasing spectral resolution and increasing the number of data points each resonance spans, it may be possible to limit the artifacts discussed in chapter 5 as a result of spectral symmetrization.

Chapter 7

Conclusions

This thesis has investigated the applications of 2D JRES data to the field of metabolomics. As stated in section 1.2, the aims of this thesis were split into three sub-objectives:

1. Spectral robustness.

2. Evaluation of JRES spectra to fingerprinting approaches.

3. Ease of feature extraction.

By examining these three areas, the plan was to establish if the JRES experiment can represent data accurately enough to ensure that meaningful and useful information can be extracted from biological samples. By evaluating the data via the common, well used 'fingerprinting' approaches, 2D JRES data was be compared and contrasted to other data sources. The feature extraction objective examined the possibility of JRES spectra providing types of information that are different from those obtained using traditional 1D methods.

Chapter 2 addresses the spectral robustness of metabolomics data. Here, by examining technical replicate spectra of 1D, pJRES and 2D JRES experiments it was determined that whilst the JRES experiment produces less robust spectra than the 1D experiment (from a technical perspective), this technical variation was still less than the metabolic

variation arising from biological sources. Therefore, it can be concluded that both pJRES and JRES spectra are suitable tools for metabolomic experiments.

To compare and contrast the different experiment types, spectral Relative Standard Deviation (RSD) was employed. Alongside the examination of the JRES spectra, other spectra were also investigated. Through this analysis of different experiments, samples and preparation types, the technique of RSD was shown to be a useful and versatile tool. This is as RSD can provide quick comparisons between data sets that would otherwise be too difficult to attempt. Further to this, the creation of a library of a variety of RSD benchmark values would be of immense practical benefit to encourage and promote the use and integration of different experiments and samples from any source.

In chapter 3, the effects of variance scaling transformations were investigated and discussed. By testing pJRES and JRES spectra alongside traditional 1D experiments, this then addressed the second aim of this thesis. Here, the effects of three variance scaling methods were evaluated on the construction of simple sample classifiers created using principal component analysis on three disparate datasets. It was shown that improvements in sample classification occurred for most experiments and data sets, however, it was also shown that the log based transformations (i.e. the glog and its extension) achieved equal or better classification accuracy than both the unscaled and other scaling methods. Clearly, the evaluation of these commonly used data processing tools has many potential benefits for the metabolomics community as many experimental techniques rely heavily upon the correct classification of the biological samples for their success.

Before investigating the effectiveness of JRES spectra when using feature extraction, the spectral processing method of the experiment was investigated. Here, methods were assessed on how well they produced the most robust spectral peaks. This is an important requirement for metabolite identification and quantification and so was an essential part of investigating the third aim of this thesis. Chapter 4 describes this investigation, which

focuses upon the effects of window functions on spectral intensity and resolution in the indirect dimension. This chapter recommends the use of larger numbers of increments when acquiring spectra, along with using the processing steps of the SEM window function and skyline projection methods, so as to improve signal to noise ratios and spectral robustness. These improvements can be applied to any investigation using the JRES experiment and hence offer benefit to many NMR practitioners.

Chapter 5 describes the examination of the 2D JRES spectral line-shapes and processing methods necessary to provide an understanding of the experiment. Here, areas of potential sources of error were examined prior to the investigation into spectral deconvolution and quantification of the intact 2D JRES spectra. Specifically, the mathematical expression describing the line-shape of the 2D JRES NMR resonance was derived under relevant experimental conditions including the different apodization functions used in this thesis and at each stage of the JRES specific processing. By analysing the specific shapes of each resonance and the manner in which the spectrum is constructed, it becomes more simple to design a peak deconvolution tool with minimal error. Specifically, by examining the long tails present in the magnitude mode spectra, it is clear that any spectral deconvolution must account for any interference arising from closely spaced peaks. The investigation into the JRES specific processing also produced interesting results. In particular, the result that symmetrising a spectrum alters both the spectral profile of multiplet (by creating low intensity features between the peaks) and signal intensity has further ramifications for feature extraction. Fortunately, the potential quantification errors remain relatively minor in this case.

The issue of feature extraction of 2D JRES spectra is explored during the investigation into the quantification and spectral deconvolution of spectra presented in chapter 6. Here, the quantification errors arising from closely spaced peaks were investigated, then an algorithm for peak quantification was proposed and tested. Here it was found that whilst 2D JRES spectra have many issues that must be investigated and resolved

before accurate quantification for all metabolites may be achieved, the experiment itself can produce highly accurate results - such as 2% absolute error when compared to the gravimentrically derived quantity. Unfortunately, many issues have also been highlighted from this investigation and it is unclear how the some of the larger absolute errors arose. This thesis demonstrates that large errors are produced when peaks are closely spaced (less than 5 full width half maxima). However, this is accounted for with spectral processing choices aimed at minimising this effect, as well as using peak structures rather than single peaks in the quantification algorithm. Spectral congestion and peak identification also highlighted as known sources of error during the testing of the algorithm on a series of chemically defined samples. Interestingly, when comparing the quantification estimates acquired from the line-shape method using 2D JRES spectra to estimates acquired using commercially available software (NMR Suite by Chenomx) and 1D spectra, the results were very similar for many samples.

Although it is clear that whilst further investigations are needed, the potential for quantitative results from the JRES experiment has been demonstrated by this thesis. It has been shown that JRES spectra have the potential to extract useful information from biological samples using established techniques - sometimes producing better results than their 1D counterparts. It is then concluded that the JRES experiment has a clear place in the toolkit of future metabolomic investigations.

7.1 Further work

Clearly, whilst this thesis has provided an in depth analysis and review of 2D JRES metabolomics, there are issues that are still to be investigated and resolved. Firstly, as described in section 6.3, there are many issues that must be addressed before 2D JRES spectra can be routinely used for semi-automated metabolite identification and quantification. Of obvious importance, is the issue of metabolite identification; without confidence in determining if a peak is correctly classified as the metabolite of interest,

the output of the peak picking algorithm is subject to uncertainty. Investigation into optimum spectral resolution of the JRES experiment may also be warranted, as in this work, it is clear that higher resolution in the indirect dimension aids peak deconvolution and integration. Peak deconvolution methods are also needed, as accurately estimating overlapping peaks may reduce errors in highly congested regions.

Other potential further work includes the identification of why the glog transformation provides excellent results for some sample types, but produces poorer results for others. Hence the creation of clear guidelines upon the use of the ex-glog transformation are also needed rather than the 'trial and error' method currently employed. Finally, the establishment of an RSD library would also be of great use to the metabolomics community (as described above), but careful consideration of how, where and what is needed to be stored needs careful consideration.

List of references

[1] M. Ala-Korpela, Y. Hiltunen, J. Jokisaari, S. Eskelinen, K. Kiviniitty, M. J. Savolainen, and Y. A. Kesniemi, *A comparative study of ^1H NMR lineshape fitting analyses and biochemical lipid analyses of the lipoprotein fractions VLDL, LDL and HDL, and total human blood plasma*, NMR in Biomedicine **6** (1993), no. 3, 225–233.

[2] H. Antti, T.M.D. Ebbels, H.C. Keun, M.E. Bollard, O. Beckonert, J.C. Lindon, J.K. Nicholson, and E. Holmes, *Statistical experimental design and partial least squares regression analysis of biofluid metabonomic NMR and clinical chemistry data for screening of adverse drug effects*, Chemometrics and intelligent laboratory systems (2004), no. 73, 139–149.

[3] R. Baldock and J. Graham, *Image processing and analysis*, Oxford University Press, 2000.

[4] J. Cavanagh, *Protein NMR spectroscopy, principles and practice*, Academic Press, 1996.

[5] D.P. Cherney, D.R. Ekman, D.J. Dix, and T.W. Collette, *Raman spectroscopy-based metabolomics for differentiating exposures to triazole fungicides using rat urine*, Analytical Chemistry **79** (2007), no. 19, 7324–7332.

[6] W.B. Dunn, N.J.C. Bailey, and H.E. Johnson, *Measuring the metabolome: current analytical technologies*, Analyst **130** (2005), 606–625.

[7] W.B. Dunn and D.I. Ellis, *Metabolomics: Current analytical platforms and methodologies*, Trends in Analytical Chemistry **24** (2005), no. 4, 285–294.

[8] T.M.D. Ebbels, E. Holmes, J.C. Lindon, and J.K. Nicholson, *Evaluation of metabolic variation in normal rat strains from a statistical analysis of 1H NMR spectra of urine*, Journal of Pharmaceutical and Biomedical Analysis **36** (2004), no. 4, 823 – 833.

[9] D.R. Ekman, Q. Teng, K.M. Jensen, D. Martinovic, D.L. Villeneuve, G.T. Ankley, and T.W. Collette, *NMR analysis of male fathead minnow urinary metabolites: A potential approach for studying impacts of chemical exposures*, Aquatic Toxicology **85** (2007), no. 2, 104 – 112.

[10] T.W.M. Fan, *Metabolite profiling by one- and two-dimensional NMR analysis of complex mixtures*, Progress in Nuclear Magnetic Resonance Spectroscopy (1996), 161 – 219.

[11] J. Förster, I. Famili, P. Fu, B. Palsson, and J. Nielsen, *Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network*, Genome Research **13** (2003), 244 – 253.

[12] P.J.D. Foxall, J.A. Parkinson, I.H. Sadler, J.C. Lindon, and J.K. Nicholson, *Analysis of biological fluids using 600 MHz proton NMR spectroscopy: Application of homonuclear two-dimensional J-resolved spectroscopy to urine and blood plasma for spectral simplification and assignment*, Journal of Pharmaceutical and Biomedical Analysis **11** (1993), no. 1, 21 – 31.

[13] P.J.D. Foxall, M. Spraul, R.D. Farrant, L.C. Lindon, G.H. Neild, and J.K. Nicholson, *750 MHz 1H-NMR spectroscopy of human blood plasma*, Journal of Pharmaceutical and Biomedical Analysis **11** (1993), no. 4-5, 267 – 276.

[14] L. Gengying and X. Haibin, *Digital quadrature detection in nuclear magnetic resonance spectroscopy*, Review of scientific instruments **70** (1999), no. 2, 1511–1513.

[15] M. Goldman, *Quantum description of high-resolution NMR in liquids*, 1992, Oxford University Press.

[16] S. Golotvin and A. Williams, *Improved baseline recognition and modeling of FT NMR spectra*, Journal of Magnetic Resonance **146** (2000), 122–125.

[17] R. Goodacre, S. Vaidyanathan, W.B. Dunn, G.G. Harrigan, and D.B. Kell, *Metabolomics by numbers: acquiring and understanding global metabolite data*, Trends in Biotechnology **22** (2004), no. 5, 245 – 252.

[18] H. Grage and M. Akke, *A statistical analysis of NMR spectrometer noise*, Journal of magnetic resonance **162** (2003), 176–188.

[19] F. Guenneau, P. Mutzenhardt, D. Grandclaude, and D. Canet, *Measurement of longitudinal and rotating frame relaxation times through fully J-decoupled homonuclear spectra*, Journal of Magnetic Resonance **140** (1999), no. 1, 250 – 258.

[20] U.L. Günther, C. Ludwig, and H. Rüterjans, *NMRLAB–advanced NMR data processing in matlab*, Journal of Magnetic Resonance **145** (2000), no. 2, 201 – 208.

[21] Y. Hiltunen, M. Ala-Korpela, J. Jokisaari, S. Eskelinen, K. Kiviniitty, M. Savolainen, and Y. A. Kesniemi, *A lineshape fitting model for 1H NMR spectra of human blood plasma*, Magnetic Resonance in Medicine **21** (1991), 222–232.

[22] A. Hines, G.S. Oladiran, J.P. Bignell, G.D. Stentiford, and M.R. Viant, *Direct sampling of organisms from the field and knowledge of their phenotype: Key recommendations for environmental metabolomics*, Environmental Science and Technology 2007 **41** (2007), no. 9, 3375 – 3381.

[23] P.J. Hore, *Nuclear magnetic resonance*, Oxford University Press, 1995.

[24] H. Hu, Q.N. Van, V.A. Mandelshtam, and A.J. Shaka, *Reference deconvolution, phase correction, and line listing of NMR spectra by the 1D filter diagonalization method*, Journal of Magnetic Resonance **134** (1998), no. 1, 76 – 87.

[25] R. Huo, R. Wehrens, and L. M. C. Buydens, *Improved DOSY NMR data processing by data enhancement and combination of multivariate curve resolution with non-linear least square fitting*, Journal of Magnetic Resonance **169** (2004), no. 2, 257 – 269.

[26] J.J. Jansen, H.C.J. Hoefsloot, H.F.M. Boelens, J. van der Greef, and A.K. Smilde, *Analysis of longitudinal metabolomics data*, Bioinformatics **20** (2004), no. 15, 2438 2446.

[27] D. Jeannerat and G. Bodenhausen, *Determination of coupling constants by deconvolution of multiplets in NMR*, Journal of Magnetic Resonance **141** (1999), no. 1, 133 – 140.

[28] I.T. Jolliffe, *Principal component analysis*, 2nd ed., Springer, 2004.

[29] M. Katajamaa and M. Oresic, *Data processing for mass spectrometry-based metabolomics*, Journal of Chromatography A **1158** (2007), no. 1-2, 318 – 328, Data Analysis in Chromatography.

[30] K. Kelley, *Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach*, Behavior Research Methods **39** (2007), no. 4, 755–766.

[31] H. Keun, T. Ebbels, H. Antti, M. Bollard, O. Beckonert, E. Holmes, J. Lindon, and J. Nicholson, *Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling*, Analytica chimica acta **490** (2003), 265 – 276.

[32] H.C. Keun, *Metabonomic modeling of drug toxicity*, Pharmacology and Therapeutics **109** (2006), no. 1-2, 92 – 106.

[33] H.C. Keun, T.M.D. Ebbels, H. Antti, M.E. Bollard, O. Beckonert, G. Schlotterbeck, H. Senn, U. Niederhauser, E. Holmes, J.C. Lindon, and J.K. Nicholson, *Analytical reproducibility in 1H NMR-based metabonomic urinalysis*, Chemical Research in Toxicology **15** (2002), no. 11, 1380–1386.

[34] T. Lange, R.F. Schulte, and P. Boesiger, *Quantitative j-resolved prostate spectroscopy using two-dimensional prior-knowledge fitting*, Magnetic Resonance in Medicine **59** (2008), 966 – 972.

[35] G.C. Lee and D.L. Woodruff, *Beam search for peak alignment of NMR signals*, Analytica Chimica Acta **513** (2004), no. 2, 413 – 416.

[36] C. Y. Lin, H. Wu, R.S. Tjeerdema, and M.R. Viant, *Evaluation of metabolite extraction strategies from tissue samples using NMR metabolomics*, Metabolomics **3** (2007), no. 1, 1573 – 3890.

[37] J.C. Lindon, E. Holmes, and J.K. Nicholson, *Pattern recognition methods and applications in biomedical magnetic resonance*, Progress in Nuclear Magnetic Resonance Spectroscopy **39** (2001), no. 1, 1 – 40.

[38] ———, *Metabonomics techniques and applications to pharmaceutical research and development*, Pharmaceutical Research **23** (2006), no. 6, 1075 – 1088.

[39] J.C. Lindon, J.K. Nicholson, E.e Holmes, and J.R. Everett, *Metabonomics: Metabolic processes studied by NMR spectroscopy of biofluids*, Concepts in Magnetic Resonance **12** (2000), no. 5, 289–320.

[40] D. MacDougall, F.J. Amore, G.V. Cox, D.G. Crosby, F.L. Estes, D.H. Freeman, W.E. Gibbs, G.E. Gordon, L.H. Keith, J. Lal, R.R. Langer, N.I. McClelland, W.F. Phillips, R.B. Pojasek, R.E. Sievers, R.G. Smerko, D.C. Wimert, W.B. Crummett, R. Libby, H.A. Laitinen, M.M. Reddy, and J.K. Taylor, *Guidelines for data acquisition and data quality evaluation in environmental chemistry*, Analytical Chemistry **52** (1980), no. 14, 2242–2249.

[41] R.P. Maharjan and T. Ferenci, *Global metabolite analysis: the influence of extraction methodology on metabolome profiles of escherichia coli*, Analytical Biochemistry **313** (2003), no. 1, 145 – 154.

[42] S.H. Moolenaar, M.S. van der Knaap, U.F.H. Engelke, P.J.W. Pouwels, F.S.M. Janssen-Zijlstra, N.M. Verhoeven, C. Jakobs, and R.A. Wevers, *In vivo and in vitro NMR spectroscopy reveal a putative novel inborn error involving polyol metabolism*, NMR in Biomedicine **14** (2001), 167 – 176.

[43] G.J. Moore and L.O. Sillerud, *The pH dependence of chemical shift and spin-spin coupling for citrate*, Journal of Magnetic Resonance, Series B **103** (1994), no. 1, 87 – 88.

[44] G.A. Morris, H. Barjat, and T.J. Home, *Reference deconvolution methods*, Progress in Nuclear Magnetic Resonance Spectroscopy **31** (1997), no. 2-3, 197 – 257.

[45] N. Morrison, D. Bearden, J.G. Bundy, T. Collette, F. Currie, M.P. Davey, N.S. Haigh, D. Hancock, O.A.H. Jones, S. Rochfort, S.A. Sansone, D. Štys, Q. Teng, D. Field, and M.R. Viant, *Standard reporting requirements for biological samples in metabolomics experiments: environmental context*, Metabolomics **3** (2007), no. 3, 203–210.

[46] H.M. Parsons, D.R. Ekman, T.W. Collette, and M.R. Viant, *Spectral relative standard deviation: a practical benchmark in metabolomics*, The Analyst (2009), 478 – 485.

[47] H.M. Parsons, C. Ludwig, U. Günter, and M.R. Viant, *Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation*, BMC Bioinformatics **8** (2007), 234.

[48] H.M. Parsons, C. Ludwig, and M.R. Viant, *Line-shape analysis of j-resolved NMR spectra: application to metabolomics and quantification of intensity errors from signal processing and high signal congestion*, Journal of Magnetic Resonance in Chemistry (2009).

[49] A.J. Pell, R.A.E. Edden, and J. Keeler, *Broadband proton-decoupled proton spectra*, Magnetic Resonance in Chemistry **45** (2007), no. 4, 296–316.

[50] A.J. Pell and J. Keeler, *Two-dimensional J-spectra with absorption-mode lineshapes*, Journal of Magnetic Resonance **189** (2007), no. 2, 293 – 299.

[51] P.V. Purohit, D.M. Rocke, M.R. Viant, and D.L. Woodruff, *Discrimination models using variance-stabilizing transformation of metabolomic NMR data*, OMICS **8** (2004), no. 2, 118 – 130.

[52] G. Reynolds, M. Wilson, A. Peet, and T.N. Arvanitis, *An algorithm for the automated quantitation of metabolites in in vitro nmr signals*, Magnetic Resonance in Medicine **56** (2006), 1211 – 1219.

[53] B.D. Ripley, *Pattern recognition and neural networks*, Cambridge University Press, 2001.

[54] D.M. Rocke and B. Durbin, *Approximate variance-stabilizing transformations for gene-expression microarray data*, Bioinformatics **19** (2003), no. 8, 966–972.

[55] _____, *Estimation of transformation parameters for microarray data*, Bioinformatics **19** (2003), no. 11, 1360–1367.

[56] D.M. Rocke and S. Lorenzato, *A two-component model for measurement error in analytical chemistry*, Technometrics **37** (1995), no. 2, 176 – 184.

[57] L.M. Samuelsson, L. Forlin, G. Karlsson, M. Adolfsson-Erici, and D.G.J. Larsson, *Using NMR metabolomics to identify responses of an environmental estrogen in blood plasma of fish*, Aquatic Toxicology **78** (2006), no. 4, 341 – 349.

[58] C.W. Schmidt, *Metabolomics: what's happening downstream of DNA*, Environmental Health Perspectives **112** (2004), no. 7, A410A415.

[59] Y.I. Shurubor, U. Paolucci, B.F. Krasnikov, W.R. Matson, and B.S. Kristal, *Analytical precision, biological variation, and mathematical normalization in high data density metabolomics*, Metabolomics **1** (2005), no. 1, 75–85.

[60] B.W. Silverman, *Density estimation for statistics and data analysis*, Chapman and Hall, 1998.

[61] A.D. Southam, T.G. Payne, H.J. Cooper, T.N. Arvanitis, and M.R. Viant, *Dynamic range and mass accuracy of wide-scan direct infusion nanoelectrospray fourier transform ion cyclotron resonance mass spectrometry-based metabolomics increased by the spectral stitching method*, Analytical Chemistry **79** (2007), no. 12, 4595–4602.

[62] H. Tang, Y. Wang, J.K. Nicholson, and J.C. Lindon, *Use of relaxation-edited one-dimensional and two dimensional nuclear magnetic resonance spectroscopy to improve detection of small metabolites in blood plasma*, Analytical Biochemistry **325** (2004), 260–272.

[63] O. Teahan, S. Gamble, E. Holmes, J. Waxman, J.K. Nicholson, C. Bevan, and H.C. Keun, *Impact of analytical bias in metabonomic studies of human blood serum and plasma*, Analytical Chemistry **78** (2006), no. 13, 4307–4318.

[64] S. Tiziani, A. Lodi, F.L. Khanim, M.R. Viant, C.M. Bunce, and U.L. Günter, *Metabolomic profiling of drug responses in acute myeloid leukaemia cell lines*, PLoS One **4** (2009), no. 1, e4251.

[65] S. Tiziani, A. Lodi, C. Ludwig, H.M. Parsons, and M.R. Viant, *Effects of the application of different window functions and projection methods on processing of 1H J-resolved nuclear magnetic resonance spectra for metabolomics*, Analytica Chimica Acta **610** (2008), no. 1, 80 – 88.

[66] N. Trbovic, F. Dancea, T. Langer, and U. Günther, *Using wavelet de-noised spectra in NMR screening*, Journal of Magnetic Resonance **173** (2005), 280 287.

[67] W. Tuffnail, G.A. Mills, P. Cary, and R. Greenwood, *An environmental ^1H NMR metabolomic study of the exposure of the marine mussel Mytilus edulis to atrazine, lindane, hypoxia and starvation*, Metabolomics **5** (2009), no. 1, 33–43.

[68] Q.H. Van and A.J. Shaka, *Improved cross peak detection in two-dimensional proton NMR spectra using excitation sculpting*, Journal of Magnetic Resonance **132** (1998), no. 1, 154 – 158.

[69] R. van den Berg, H. Hoefsloot, J. Westerhuis, A. Smilde, and M. van der Werf, *Centering, scaling, and transformations: improving the biological information content of metabolomics data*, BMC Genomics **7** (2006), no. 1, 142.

[70] K. E. van Holde, W. C. Johnson, and P. S. Ho, *Principles of physical biochemistry*, 2nd ed., Pearson Prentice Hall, 1998.

[71] M.R. Viant, *Improved methods for the acquisition and interpretation of NMR metabolomic data*, Biochemical and Biophysical Research Communications **310** (2003), 943 – 948.

[72] M.R. Viant, C. Ludwig, S. Rhodes, U.L. Günther, and D. Allaway, *Validation of a urine metabolome fingerprint in dog for phenotypic classification*, Metabolomics **3** (2007), no. 4, 453 – 463.

[73] M.R. Viant, B.G. Lyeth, M.G. Miller, and R.F. Berman, *An NMR metabolomic investigation of early metabolic disturbances following traumatic brain injury in a mammalian model*, NMR in Biomedicine **18** (2005), no. 8, 507–516.

[74] M.R. Viant, E.S. Rosenblum, and R.S. Tjeerdema, *NMR-based metabolomics: A powerful approach for characterizing the effects of environmental stressors on organism health*, Environmental Science and Technology **37** (2003), 4982 – 4989.

[75] Y. Wang, M.E. Bollard, H. Keun, H. Antti, O. Beckonert, T.M. Ebbels, J.C. Lindon, E. Holmes, H. Tang, and J.K. Nicholson, *Spectral editing and pattern recognition methods applied to high-resolution magic-angle spinning 1H nuclear magnetic resonance spectroscopy of liver tissues*, Analytical Biochemistry **323** (2003), no. 1, 26 – 32.

[76] A. Webb, *Statistical pattern recognition*, 2nd ed., John Wiley and sons, 2002.

[77] W. Weckwerth, *Metabolomics in systems biology*, Annual Review of plant biology **54** (2003), 669–89.

[78] A.M. Weljie, J. Newton, P. Mercier, E. Carlson, and C.M. Slupsky, *Targeted profiling: Quantitative analysis of ^1H NMR metabolomics data*, Analytical Chemisty (2006), no. 78, 4430–4442.

[79] H.T. Widarto, E Van Der Meijden, A.W.M. Lefeber, C. Erkelens, H.K. Kim, Y.H. Choi, and R Verpoorte, *Metabolomic differentiation of brassica rapa following herbivory by different insect instars using two-dimensional nuclear magnetic resonance spectroscopy*, Journal of chemical ecology **32** (2006), no. 11, 2417–2428.

[80] D. H. Williams and I. Flemming, *Spectroscopic methods in organic chemistry*, 5th ed., McGraw-Hill, 1995.

[81] H. Wu, A.D. Southam A., Hines, and M.R. Viant, *High-throughput tissue extraction protocol for NMR- and MS-based metabolomics*, Analytical Biochemistry **372** (2008), 204–212.

[82] Y. Xi, J.S. de Ropp, M.R. Viant, D.L. Woodruff, and P. Yu, *Automated screening for metabolites in complex mixtures using 2D COSY NMR spectroscopy*, Metabolomics **2** (2006), no. 4, 221–233.

[83] Y. Xi and D. Rocke, *Baseline correction for NMR spectroscopic metabolomics data analysis*, BMC Bioinformatics **9** (2008), no. 1, 324.

Appendix A

Parameter Values

The following table lists all the parameter values of the glog transformations and spectral simulations used in the investigations discussed in this thesis.

Data set and experiment type	Generalised log	Extended glog
1D mussel	$\lambda = 2.0025 \times 10^{-8}$	$\lambda = 1.2689 \times 10^{-8}$ $y_0 = 8.7026 \times 10^{-5}$
1D canine	$\lambda = 6.8457 \times 10^{-7}$	$\lambda = 2.3969 \times 10^{-7}$ $y_0 = 3.5655 \times 10^{-4}$
pJRES canine	$\lambda = 2.3024 \times 10^{-9}$	$\lambda = 1.5175 \times 10^{-9}$ $y_0 = 4.9506 \times 10^{-5}$
JRES canine	$\lambda = 1.1802 \times 10^{-13}$	$\lambda = 4.0877 \times 10^{-12}$ $y_0 = 1.6196 \times 10^{-6}$
1D Flounder	$\lambda = 3.1928 \times 10^{-7}$	$\lambda = 6.8641 \times 10^{-10}$ $y_0 = 2.3017 \times 10^{-5}$
pJRES Flounder	$\lambda = 1.262 \times 10^{-8}$	$\lambda = 2.3888 \times 10^{-9}$ $y_0 = 7.944 \times 10^{-5}$
JRES Flounder	$\lambda = 6.9974 \times 10^{-14}$	$\lambda = 4.0877 \times 10^{-12}$ $y_0 = 1.575 \times 10^{-6}$

Table A.1: Parameter values for all glog transformations constructed in chapter 3.3.

Parameter	Name	Value
T_2	Effective transverse relaxation time (direct dimension)	0.679s
τ_2	Effective transverse relaxation time (indirect dimension)	4.125s
a_1, a_2	Acquisition time	6.239s
L	Line broadening	0.5Hz

Table A.2: Spectral parameters for simulated NMR line-shapes.

Appendix B

Additional statistics

This appendix includes additional statistics from chapters 2 and 3 which are not discussed in main text.

Experiment	Scaling	Cross validation accuracy
1D mussel	Unscaled	37.04%
	Autoscaled	33.33%
	Pareto scaled	51.85%
	Glog transform	100.00%
	Ex-glog transform	96.30%
1D canine	Unscaled	38.89%
	Autoscaled	80.56%
	Pareto scaled	83.33%
	Glog transform	83.33%
	Ex-glog transform	83.33%
pJRES canine	Unscaled	32.43%
	Autoscaled	83.78%
	Pareto scaled	56.78%
	Glog transform	83.78%
	Ex-glog transform	83.33%
JRES canine	Unscaled	40.54%
	Autoscaled	56.76%
	Pareto scaled	56.76%
	Glog transform	83.78%
	Ex-glog transform	83.78%
1D flounder	Unscaled	86.84%
	Autoscaled	86.84%
	Pareto scaled	86.84%
	Glog transform	92.11%
	Ex-glog transform	73.68%
pJRES flounder	Unscaled	76.32%
	Autoscaled	81.58%
	Pareto scaled	76.32%
	Glog transform	86.84%
	Ex-glog transform	84.21%
JRES flounder	Unscaled	68.42%
	Autoscaled	63.16%
	Pareto scaled	86.84%
	Glog transform	86.84%
	Ex-glog transform	100.00%

Table B.1: Leave-one-out cross-validation statistics for each of the PCA-LDA models constructed in section 3.3

Species	Sample	Class	Extraction method	Analytical technique	Variation across technical replicates		Inter-individual variation	
					Median RSD	RSD range	Median RSD	RSD range
Chub	liver	-	M:W	1D NMR	4.6%	53%		
	liver	-	Perchloric acid	1D NMR	20.6%	267%		
3-spined stickleback	liver	-	M:C:W	1D NMR	3.4%	32%		
European flounder	liver	-	M:C:W	1D NMR	3.1%	34%		
	liver	-	M:C:W	pJRES NMR	12.5%	58%		
	liver	-	M:C:W	JRES NMR	19.8%	97%		
	liver	field (Alde)	M:C:W	1D NMR			30.1%	183%
	liver	field (Tyne)	M:C:W	1D NMR			24.9%	198%
Dab	liver	-	M:C:W	FT-ICR MS	13.1%	68%		
fathead minnow	testis[a]	control	M:C:W	1D NMR			29.4%[b]	122%
	plasma[a]	control	pH 7.4 buffer	1D NMR			58.4%	174%
	urine[a]	control	pH 7.4 buffer	1D NMR			52.9%	146%

Table B.2: Summary of spectral RSDs for multiple fish metabolomics datasets.
[a] All samples obtained from the same fathead minnows.
[b] To confirm that spectral processing has a minimal effect on the median RSD, this dataset (originally apodised, Fourier transformed, phased, baseline corrected and calibrated using ACD/1D NMR Processor software) was reprocessed using Topspin software, yielding a median RSD of 26.7%.

Species	Sample	Class	Extraction method	Analytical technique	Variation across technical replicates		Inter-individual variation	
					Median RSD	RSD range	Median RSD	RSD range
Red abalone	foot muscle	-	Perchloric acid	1D NMR	5.4%	179%		
	foot muscle[c]	healthy	Perchloric acid	1D NMR			16.0%	125%
	foot muscle	diseased	Perchloric acid	1D NMR			21.5%	146%
	digestive gland[c]	healthy	Perchloric acid	1D NMR			19.7%	132%
	haemolymph[c]	healthy	Perchloric acid	1D NMR			25.2%	92%
Marine mussel	muscle	-	M:C:W	1D NMR	6.1%	56%		
	muscle	field	M:C:W	1D NMR			24.4%	111%
	muscle	laboratory	M:C:W	1D NMR			26.0%	138%
	mantle	field	M:C:W	1D NMR			23.5%	90%
	mantle	laboratory	M:C:W	1D NMR			26.9%	97%
Dog	urine	-	buffer to pH 7.05 ± 0.05	1D NMR	1.6%	51%		
Rat	brain[d]	control	Perchloric acid	1D NMR			7.2%	87%
	plasma[d]	control	pH 7.4 buffer added	1D NMR			8.0%	106%
	urine[d]	control	pH 7.4 buffer added	1D NMR			32.2%	164%
K562 cell line	cell extract	-	M:C:W	1D NMR	14.0%	90%		
	cell extract	untreated	M:C:W	1D NMR			20.5%	84%
	cell extract	treated	M:C:W	1D NMR			22.0%	112%

Table B.3: Summary of spectral RSDs for multiple marine invertebrates and mammalian metabolomics datasets

Appendix C

Gaussian and Rayleigh distributions

Consider two Gaussian distributed variables x and y with equal variance, σ and zero mean. The probability density function of each variable is then given as:

$$\int \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-x^2}{2\sigma^2}\right) \mathrm{d}x \tag{C.1}$$

Squaring and adding the variables then gives a density function of

$$\int\int \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-x^2}{2\sigma^2}\right)\right]^2 + \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-y^2}{2\sigma^2}\right)\right]^2 \mathrm{d}x\mathrm{d}y = \int\int \frac{1}{2\pi\sigma^2} \exp\left(\frac{-2(x^2+y^2)}{4\sigma^2}\right) \mathrm{d}x\mathrm{d}y \tag{C.2}$$

Changing to polar coordinates, $(x, y) \rightarrow (r\cos\theta, r\sin\theta)$, then gives equation C.2 as:

$$\int\int \left[\frac{1}{2\pi\sigma^2} \exp\left(\frac{-r^2}{2\sigma^2}\right)\right] \left|\frac{\partial(x,y)}{\partial(r,\theta)}\right| \mathrm{d}r\mathrm{d}\theta \tag{C.3}$$

In this case, the Jacobian, $|\frac{\partial(x,y)}{\partial(r,\theta)}|$, is equal to r. Also noting that there is no dependacy upon θ in the integrand then yields

$$\frac{1}{2\pi} \int \left[\frac{r}{\sigma^2} \exp\left(\frac{-r^2}{2\sigma^2}\right)\right] \mathrm{d}r \tag{C.4}$$

Which is a Rayleigh distribution. □