University of London
Imperial College of Science, Technology and Medicine
Department of Epidemiology & Biostatistics

Statistical Methods in Metabolomics

Harriet Jane Muncey

Supervised by Dr Maria De Iorio & Dr Tim Ebbels

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Epidemiology & Biostatistics of the University of London and
the Diploma of Imperial College, October 16, 2014

Declaration

I herewith certify that all material in this dissertation which is not my own work has been properly acknowledged.

Harriet Muncey

Abstract

Metabolomics lies at the fulcrum of the system biology 'omics'. Metabolic profiling offers researchers new insight into genetic and environmental interactions, responses to pathophysiological stimuli and novel biomarker discovery. Metabolomics lacks the simplicity of a single data capturing technique; instead, increasingly sophisticated multivariate statistical techniques are required to tease out useful metabolic features from various complex datasets. In this work, two major metabolomics methods are examined: Nuclear Magnetic Resonance (NMR) Spectroscopy and Liquid Chromatography-Mass Spectrometry (LC-MS). MetAssimulo, an ^{1}H-NMR metabolic-profile simulator, was developed in part by this author and is described in the Chapter 2. Peak positional variation is a phenomenon occurring in NMR spectra that complicates metabolomic analysis so Chapter 3 focuses on modelling the effect of pH on peak position. Analysis of LC-MS data is somewhat more complex given its 2-D structure, so I review existing pre-processing and feature detection techniques in Chapter 4 and then attempt to tackle the issue from a Bayesian viewpoint. A Bayesian Partition Model is developed to distinguish chromatographic peaks representing useful features from chemical and instrumental interference and noise. Another of the LC-MS pre-processing problems, data binning, is also explored as part of H-MS: a pre-processing algorithm incorporating wavelet smoothing and novel Gaussian and Exponentially Modified Gaussian peak detection. The performance of H-MS is compared alongside two existing pre-processing packages: apLC-MS and XCMS.

Dedication

For Silvia; who showed me my strength.

Acknowledgements

I would like to thank Dr Maria De Iorio and Dr Timothy Ebbels for their invaluable expert guidance and patience during my PhD research. I would like to acknowledge Paul Benton, Imperial College London, for his expert advice in getting to grips with XCMS and Will Astle for his patient help in understanding NMR data.

I also thank Laura Egnash and Michael Reilly, formerly of the Department of Discovery Biomarkers, Pfizer Global R&D, Ann Arbor, MI, for providing the Synthetic Urine data used in testing the H-MS algorithm in Chapter 4. Alexander Amberg of Sanofi-Aventis Deutschland GmbH, DSAR Preclinical Safety, Frankfurt, Germany, is acknowledged for providing the Standard Dilution Series data also used in the testing of H-MS. Dr Elizabeth Want of Imperial College London enabled me access to the LC-MS datasets used in Chapter 4. Gregory Tredwell, Imperial College London, kindly provided the ^{1}H-NMR data used in Chapter 3. My financial support is acknowledged from an MRC Capacity Building PhD Studentship.

I am unbelievably grateful to my friends, in particular Lizzy, Laurie and Ratchet for their unfaltering support throughout my academic career. From extreme study sessions during my undergraduate degree to random nonsense and escapades in Durham, London and many other places; they have always been there to both support and distract me as well as keep me going. Highlights include The Sandwich of Dreams, Pie, Quiche and of course Flan. I would also like to thank my family for their love and understanding, and without whom I never would have got this far. Lastly, but by no means least, I wish to thank Silvia for showing me back to my path and helping me realise it was worth it.

Contents

Abstract **iii**

Acknowledgements **vii**

1 Introduction **1**

1.1 Metabolomics 1

1.2 Technologies 2

1.2.1 ^{1}H Nuclear Magnetic Resonance Spectroscopy 3

1.2.2 Liquid Chromatography - Mass Spectrometry 9

1.3 Bayesian Inference 15

1.4 Aims 19

2 Simulation of Realistic ^{1}H-NMR Metabolic Profiles: MetAssimulo **21**

2.1 Implementation 23

2.1.1 Setting Parameters 24

2.1.2 Pre-processing 26

2.1.3 Simulating Mixture Spectra 30

2.2 Results and Discussion 31

2.3 Conclusion 33

3 Modelling pH-induced changes in ^{1}H-NMR chemical shifts 37

3.1 Background 37

3.1.1 Modelling Titration Curves 38

3.1.2 Model Selection Criteria 41

3.2 Data 46

3.3 Model Construction 48

3.3.1 Model Tuning and Convergence 48

3.3.2 Model Selection 54

3.4 Additive Constant 62

3.4.1 Model Selection 63

3.5 Random Effects Model 65

3.5.1 Model Selection 67

3.6 Conclusion 67

4 LC-MS Processing 69

4.1 Background 69

4.1.1 Wavelet Smoothing 71

4.1.2 Peak Detection 73

4.1.3 Data Binning 81

4.2 A Bayesian Partition Model for Peak Detection 82

4.2.1 Metropolis-Hastings Algorithm for Multiple Change Points 87

4.2.2 Testing and Convergence 89

4.2.3 Sensitivity Analysis 94

4.2.4 Results 95

4.3 H-MS: Wavelet Smoothing, Adaptive Binning and Peak Detection 96

4.3.1 Binning in m/z Dimension 97

4.3.2 Peak Detection in Chromatographic Dimension 98

4.3.3 Results and Discussion 102

4.3.4 Conclusion 118

5 Conclusion **121**

Appendices **126**

.1 Glossary of Abbreviations 126

.2 Log Marginal Likelihood Estimation Algorithm 127

.3 Marginal Likelihood Derivation for Bayesian Normal-Inverse-Gamma Models . . 128

.4 Code 130

Bibliography **130**

List of Tables

3.1 Number of protonation sites for each metabolite studied. Where multiple resonances are used for a single metabolite these are labeled numerically. 47

3.2 Actual pKa values vs Posterior Mean pKa Estimates and Variance from MCMC simulation using *rjags* package and MATLAB function *nlinfit* pKa Estimates with mean squared error (MSE) values. 57

3.3 Akaike Information Criterion values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Smallest values are shown in bold, indicating chosen model. 57

3.4 Bayesian Information Criterion values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Smallest values are shown in bold, indicating chosen model. 58

3.5 Deviance Information Criterion values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Smallest values are shown in bold, indicating chosen model. 58

3.6 Actual number of free parameters for appropriate model and p_D values (estimated number of effective parameters) for each model. 59

3.7 Log Pseudo Marginal Likelihood values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Largest values are shown in bold, indicating chosen model. 59

3.8 Bayes Factors calculated using approximate marginal likelihoods, and showing preference for either the 2 or 3 Site Models over the 1 Site Model. 59

3.9 Actual pK$_a$ values vs Posterior Mean pK$_a$ Estimates and Variance for model with additive constant. 62

3.10 Akaike Information Criterion values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Smallest values are shown in bold, indicating chosen model. 63

3.11 Bayesian Information Criterion values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Smallest values are shown in bold, indicating chosen model. 63

3.12 Log Pseudo Marginal Likelihood values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Largest values are shown in bold, indicating chosen model. 64

3.13 Bayes Factors calculated using approximate marginal likelihoods, and showing preference for either the 2 or 3 Site Model. 64

3.14 Deviance Information Criterion values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Smallest values are shown in bold, indicating chosen model. 64

3.15 Actual number of free parameters for appropriate model and p_D values (estimated number of effective parameters) for each model. 65

3.16 Actual pKa values vs Posterior Mean pKa Estimates and Posterior pKa Variance for Random Effects Model. 67

3.17 Deviance Information Criterion values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Smallest values are shown in bold, indicating chosen model. 67

3.18 Actual number of free parameters for appropriate model and p_D values (estimated number of effective parameters) for each model. 67

4.1 Overview of existing pre-processing algorithms for (LC-)MS data. 80

4.2 Chemical composition of synthetic urine sample and approximate concentration of analytes. Metabolites manually identified in LC-MS runs and those automatically detected by at least one algorithm indicated by 'X'. 106

4.3 continued: Chemical composition of synthetic urine sample and approximate concentration of analytes. Metabolites manually identified in LC-MS runs and those automatically detected by at least one algorithm indicated by 'X'. 107

4.4 Chemical composition of standards dilution series batch. 'X' indicates detected in any form (including dimers and adducts) in any replicate by any of the algorithms, with the detected forms also listed. 108

4.5 Mean and variance of number of peaks detected by apLC-MS over 10 replicates of noise datasets simulated using five different Poisson process rates, for a range of values for parameters min.run (minimum length of elution time for a series of grouped signals to be considered a peak), min.pres (minimum proportion of presence in the time period for a series of signals grouped by m/z to be considered a peak) and tol (m/z tolerance level for the grouping of data points expressed as a fraction of the m/z value). 109

4.6 Mean and variance of number of peaks detected by matchedFilter over 10 'shuffled' datasets for a range of values for parameters fwhm (full width at half maximum of Rt peak model), max (maximum number of peaks per Extracted Ion Chromatogram) and snthresh (signal-to-noise threshold). 110

4.7 Mean and variance of no. of peaks detected by apLC-MS over 10 'shuffled' datasets for a range of parameters: min.run (min. length of Rt for a series of grouped signals to be considered a peak), min.pres (min. proportion of presence in the time period for a series of signals grouped by m/z to be considered a peak) and tol (m/z tolerance for the grouping of data points expressed as a fraction of the m/z value). 110

4.8 Average total number of detected and average number of matched peaks for centWave and H-MS algorithms for each dilution level. 115

4.9 R^2 for linear model fitted to dilution curves of peaks detected by both H-MS and centWave across all dilution levels. 117

4.10 R^2 for linear model fitted to dilution curves of peaks detected by H-MS but not centWave across all dilution levels. 118

4.11 R^2 for linear model fitted to dilution curves of peaks detected by centWave but not H-MS across all dilution levels. 118

List of Figures

1.1 Example of a NMR multiplet signal resulting from spin-spin coupling. 4

1.2 A typical ^{1}H NMR spectrum of a human urine sample with some labelled metabolite resonances, where the x-axis is ppm (parts per million) relative to some internal standard and the y-axis gives the signal intensity. 5

1.3 A high level view of the LC-MS workflow taken from [1]. The sample enters the liquid chromatography where compounds are separated in the time domain, then the sample is ionized and passed to the mass spectrometer where compounds are further separated by their mass. This results in a 2-dimensional dataset where signals are described by a (retention time (Rt), mass-to-charge ratio, intensity) triplet. 9

1.4 A 2-D Intensity map of LC-MS Data from [2]. x and y axes give the Retention Time (Rt) and mass-to-charge ratio (m/z) and z axis gives the intensity (ion count) detected. 12

1.5 A chromatographic peak split across three bins due to drift in m/z. 12

1.6 An isotopic cluster (proteomic data acquired on a low-resolution spectrometer) as it is visualized in a 2-D map from [3]. The x-axis and y-axis of this noise-filtered 2-D map represent the chromatography and MS dimensions, respectively. Relevant mass spectrum and mass chromatogram are represented as cross-sections. A contour plot of the 2-D map is shown in the inset. 13

2.1 A high level flow chart of the MetAssimulo algorithm. 23

2.2 (a) Real normal urine ^{1}H-NMR spectrum, (b) Mean of simulated normal urine ^{1}H-NMR spectra produced using MetAssimulo. 31

2.3 (a) Mean simulated normal urine ^{1}H-NMR spectrum, (b) Mean simulated ^{1}H-NMR spectrum of urine in paraquat poisoning produced using MetAssimulo. . 34

2.4 (a) PCA scores plot of the first two principal components for the simulated normal and diseased datasets show clear separation, (b) Loadings on PC1 indicating metabolite resonances that describe a large portion of the difference between the normal and diseased datasets, (c) Loadings on PC2 indicating metabolite that has a large within-group variance. 34

2.5 Simulated ^{1}H-NMR spectral peak shift for the two aromatic singlets of histidine. 35

2.6 Correlation matrix used to demonstrate MetAssimulo correlation functionality. . 35

2.7 Inter and intra-metabolite correlations: (i) Complete correlation matrix and insets (ii)-(v) showing strong negative inter-metabolite correlation between citrate and creatinine and positive between 2-oxoglutarate and creatinine (ii),(iii) and strong positive intra-metabolite correlations for creatinine (iv) and citrate (v). Colour scale indicates the level of Pearson correlation. 35

2.8 Pairwise correlation coefficients mapped as a colour code onto the mean spectrum. Correlations to (a) citrate 2.65ppm, (b) creatinine 4.08ppm, (c) 2-oxoglutarate 2.44ppm. 36

3.1 An example monoprotic titration curve modeled by the Henderson Hasselbach Equation is shown in blue. The pK_a value is 7 and the inflection point is indicated with a black cross. The acid and base asymptotic limits are given by cyan and red dashed lines respectively. 39

3.2 An example polyprotic titration curve model for a two site molecule is shown in blue. The pKa values are 3 and 8 with the inflection points indicated by black crosses. The acid/base asymptotic limits are given by cyan, green and red dashed lines. 41

3.3 Metabolite ^{1}H NMR titration curves showing varying effects of pH on several different labelled metabolite resonances. 47

3.4 Metabolite ^{1}H NMR titration curve for Imidazole 1 (a one site molecule with $pKa = 6.95$) 49

3.5 MCMC trace and density plot for pKa of Imidazole using 1 site model. 4 chains run with 100,000 iterations after a burn-in of 100,000 and thinning parameter of 10. 49

3.6 MCMC trace and density plot for acid and base limits of Imidazole using 1 site model. 4 chains run with 100,000 iterations after a burn-in of 100,000 and thinning parameter of 10. 50

3.7 MCMC trace and density plot for acid and base limits of Imidazole using 1 site model. 4 chains run with 500,000 iterations after a burn-in of 500,000 and thinning parameter of 100. 51

3.8 MCMC trace and density plot for acid and base limits of Imidazole using 1 site model with prior N(5,1)I(0,10). Multiple chains run for 100,000 iterations after a burn-in of 100,000 and thinning parameter of 10. 51

3.9 MCMC trace and density plot for acid and base limits of Imidazole using 1 site model with prior N(5,1)I(0,10). Multiple chains run for 500,000 iterations after a burn-in of 500,000 and thinning parameter of 50. 52

3.10 MCMC trace and density plot for acid and base limits of Imidazole using 1 site model with prior N(5,2)I(0,10). 4 chains run for 100,000 iterations after a burn-in of 100,000 and thinning parameter of 10. 53

3.11 MCMC trace and density plot for pka of Piperazine using 2 site model. 4 chains run for 100,000 iterations after a burn-in of 100,000 and thinning parameter of 10. 54

3.12 MCMC trace and density plot for acid and base limits of Piperazine using 2 site model. 4 chains run for 100,000 iterations after a burn-in of 100,000 and thinning parameter of 10. 55

3.13 MCMC trace and density plot for acid and base limits of Piperazine using 2 site model with stronger prior N(5,2)I(0,10). 4 chains run for 100,000 iterations after a burn-in of 100,000 and thinning parameter of 10. 56

3.14 Polyprotic model fit for 1 Site molecule Formate (top), 2 Site molecule Creatinine (middle) and 3 Site molecule Citrate (bottom). Model1, Model2 and Model3 are polyprotic models with $q = 1, 2, 3$ sites respectively shown from left to right. The data is shown in blue, the fit and 95% credible intervals are indicated by black lines. 60

3.15 nlinfit model fit for 1 Site molecule Formate (left), 2 Site molecule Creatinine (middle) and 3 Site molecule Citrate (right) with data given in blue and the fit shown in black. The fit is so close as to be barely distinguishable by eye. 60

4.1 Examples of two commonly used wavelet functions (a) Daubechies 4, (b) Symmlet 8 [4] 71

4.2 (a) Identified Peaks in m/z domain with SNR >3, (b) CWT coefficients of the signal at different scales, (c) 'Ridge Lines' constructed from joining the location of local maxima at different scales and indicate where 'important' features persist over multiple scales.[5] 75

4.3 A chromatographic peak split across three bins due to drift in m/z. 81

4.4 Plot of test data (red), posterior mean of distribution within each partition (blue), marginal posterior probability of each data point being a change point (green line) and those points with a Bayes Factor >10are represented by dots at the top of the diagram. 90

4.5 Figure shows the signal data modelled (blue scatter plot), the marginal posterior probability of each data-point being a change point (red line plot) correctly indicating the change points, trace plot of K over the run (blue line plot) and the posterior distribution of K (blue histogram). The run was 6,000 iterations long including 5,000 burn-in. Prior mean on θ was 20 and initial value of K was set to 20. Prior parameters p and q were set to 0.5 and 20 respectively and prior parameters d and s were set to 50 and 20 respectively. Despite the data being noisy, the model converges quickly to the correct answer. 91

4.6 Figure shows the signal data modelled (blue scatter plot), the marginal posterior probability of each data-point being a change point (red line plot) correctly indicating the change points, trace plot of K over the run (blue line plot) and the posterior distribution of K (blue histogram). The run was 6,000 iterations long including 5,000 burn-in. Prior mean on θ was 20 and initial value of K was set to 20. Prior parameters p and q were set to 0.5 and 20 respectively and prior parameters d and s were set to 50 and 20 respectively. Despite there being relatively few data-points, the model again converges quickly to the correct answer. 92

4.7 Figure shows the signal data modelled (blue scatter plot), the marginal posterior probability of each data-point being a change point (red line plot) correctly indicating the change points, trace plot of K over the run (blue line plot) and the posterior distribution of K (blue histogram). The run was 10,000 iterations long including 5,000 burn-in. Prior mean on θ was 20 and initial value of K was set to 20. Prior parameters p and q were set to 0.5 and 20 respectively and prior parameters d and s were set to 50 and 20 respectively. Given there was considerable noise and few data-points, the model took slightly longer to converge but still did so relatively quickly and to the correct answer. 93

4.8 Sensitivity Analysis for prior parameter (a) p and (b) q on k indicating that increasing variance for prior on k allows greater number of change points to be fitted. 94

4.9 Sensitivity Analysis for prior parameters (a) d and (b) s indicating that increases of d or s correspond to higher mean and variance allowing greater variability of intensity within each partition and so fitting fewer change points. 95

4.10 Shows the real data (red) overlaid with the posterior mean (blue) and the marginal probability of a change point (green). Blue dots indicate points with a Bayes Factor >10. The inset shows a close-up of a relatively low intensity peak. 96

4.11 An example Total Mass Spectrum (with zoomed inset) showing raw data in black and the wavelet-smoothed data in red. 98

4.12 Candidate peaks accepted by smoothing spline shape test using (a) Gaussian model and (b) Exponentially Modified Gaussian model. Blue circles indicate raw data points, red line indicates Gaussian fit, black line indicates EMG fit and green line indicates smoothing spline fit. 101

4.13 A 3-Hydroxycinnamic Acid peak intensity plot (a) of the raw data, with the colour scale indicating intensity and bin edge locations shown for H-MS (red lines) and apLC-MS (cyan lines). Also plotted are chromatograms of the H-MS bin (b) and the two apLC-MS bins (c, d). 111

4.14 A Creatine peak intensity plot (a) of the raw data, with the colour scale indicating intensity and bin edge locations shown for H-MS (red lines) and apLC-MS (cyan lines). Also plotted are chromatograms of the H-MS bin (g) and the five apLC-MS bins (b-f). 112

4.15 Raw intensity plots, with colour scale indicating intensity, of 6 detected metabolite peaks. Peak parameters shown for H-MS (red) and centWave (black) as boxes. (a) 4-Aminohippuric acid gives an example of centWave excluding high intensity data-points but H-MS captures the whole peak. For 3-Hydroxycinnamic acid in (b) we see a substantial drift in m/z value within a peak. H-MS includes all of the peak, whilst centWave detects two separate peaks. (c) Pimelic acid demonstrates where H-MS has defined a tighter area than centWave, including less noise. (d) Xanthine and (e) Riboflavin shows where H-MS excludes some relatively high intensity data-points that are captured by centWave. In (f) H-MS includes some lower intensity datapoints surrounding the Isocitric acid peak which perhaps should not be included as per the centWave parameters. 114

4.16 Intensity curves for all peaks detected by both a) H-MS and b) centWave across the entire dilution series. The legend gives the molecule names, with the number in front indicating the replicate sample (1, 2 or 3). Strong linear trends give confidence that metabolites have been accurately identified and matched. . . . 116

4.17 Intensity curves for all peaks detected by either H-MS, a) with closeup b), and centWave only, c) with closeup d), across the entire dilution series. The legend gives the molecule names, with the number in front indicating the replicate sample (1, 2 or 3). Lack of linear trend may be due to ion suppression or saturation effects. 116

4.18 Raw intensity plots, with the colour scale indicating intensity (ion count) at a particular data point, of (a) Hippuric acid-d2+H peak detected by both H-MS and centWave in replicate 1, (b) acetylcarnitine-d3+Na peak detected by centWave in replicate 1 and (c) 2-acetylcarnitine-d3+H peak detected by H-MS in replicate 2. Peak output parameters shown for H-MS (red) and centWave (black) as boxes. 117

Chapter 1

Introduction

1.1 Metabolomics

In the post-genomic era there has been a massive growth in ‘omics’ techniques investigating different levels of biological organisation. At the core of systems biology research is metabolic profiling, commonly known as metabolomics. Metabolomics focuses on high-throughput identification and quantification of metabolites, small molecules ($\leq$ 1500Da) involved in metabolism [6]. When attempting to relate genes to the overall function of a system, the metabolome (the complete set of metabolites of an organism) more closely reflects the activities of the organism at a functional level than, for example, the transcriptome [7]. Metabolic fluxes are not only regulated by gene expression, but also by additional factors, which include the abundance of metabolites as substrates (molecules acted upon by enzymes) and products (the output of a reaction or process) [8]. Therefore metabolic profiling adds another dimension to our understanding of biological systems. Figure 1.1 illustrates, somewhat simplistically, how metabolomics fits into the ‘omics’ structure of systems biology. In reality, the interactions between omics are not one-way only and much research is devoted to understanding the complex ’inter-omics’ mechanisms in order to build up a holistic view of an organism’s function. The omics research areas mentioned here are by no means an exhaustive list and the depth and breadth of omics specialities is continuously expanding.

$$\boxed{\text{Genomics}} \rightarrow \boxed{\text{Transcriptomics}} \rightarrow \boxed{\text{Proteomics}} \rightarrow \boxed{\text{Metabolomics}} \tag{1.1}$$

With the relatively recent development of systems biology comes the promise of 'personalised health-care solutions' and an improved understanding of molecular epidemiology [9]. The applications of accurate metabolic profiling reach from studying drug toxicity and pharmacology to disease-screening for conditions such as diabetes or cancer [10, 11, 12, 13, 14]. Metabolomics, as an expression of genetic and environmental factors, is key to furthering our knowledge of how humans and other organisms function as individuals and interacting complex systems.

There are generally considered around 2000 major metabolites for humans, however this number increases significantly when secondary metabolites are explored [15, 16]. Although the terms 'metabonomics', 'metabolomics', metabolic 'fingerprinting' or 'profiling' were assigned subtly different definitions originally, they are usually used interchangeably today. Metabolomics can be formally defined as 'the comprehensive quantitative analysis of all the metabolites of an organism or specified biological sample', typically involving 'the quantitative measurement of the multi-parametric time-related metabolic responses of a complex (multi-cellular) system to a pathophysiological intervention or genetic modification' [9].

1.2 Technologies

Metabolomics lacks a definitive data-acquisition technique, with a variety of methods offering different advantages and disadvantages depending on the aims or priorities of a study or experiment. However, the two main techniques for capturing metabolite data are ^{1}H Nuclear Magnetic Resonance (NMR) Spectroscopy and Mass Spectrometry (MS) [17, 18]. Metabolites in biofluids are in dynamic equilibrium with those in cells and tissues so their metabolic profile reflects changes in the state of an organism due to disease or environmental effects. NMR is highly reproducible, offering the advantage of being non-destructive, and is able to give a nearly global metabolite profile including structural information enabling the identification of the most abundant metabolites. MS boasts a significantly superior sensitivity [19, 17], but has great variability of results as consequence which is exacerbated by time-consuming and error-prone sample preparation procedures. There are efforts under way within the metabolomic community to marry the merits of both NMR and MS and forge a multi-platform approach to metabolomics research [20, 21, 22]. Whilst NMR spectra are typically characterised by a 1-dimensional signal whose intensity varies in proportion to metabolite concentrations detected across a scale of 'parts per million' (ppm) of an internal standard, MS is usually paired with a chromatographic technique such as liquid or gas chromatography (LC-MS or GC-MS [23]) resulting in a 2-dimensional dataset of intensities varying in both retention time (Rt) and along the mass-to-charge ratio scale (m/z). Metabolites present as 'peaks' in the data, where the instrument has registered the presence of a molecular species within the biofluid. NMR machines are typically more expensive with faster running time, but LC-MS machines certainly have more widespread availability. Use of GC-MS has largely been superseded by LC-MS due to more convoluted sample preparation, limitations on the size and type of molecules that can be measured as well as slow speed of data acquisition. Despite this, GC-MS is still used in plant metabolomics.

1.2.1 ^{1}H Nuclear Magnetic Resonance Spectroscopy

^{1}H-NMR spectroscopy gives an almost global metabolic profile as it has the potential to detect nearly all proton-containing metabolites and allows metabolites to be detected simultaneously without pre-selection. Biofluids contain thousands of metabolites, but a typical NMR spectrum will only contain signals from a few hundred of the most abundant metabolites. Despite relatively poor sensitivity in comparison with analytical methods such as mass spectrometry, NMR spectroscopy requires minimal sample preparation meaning the process is highly reproducible, and is able to measure concentrations as low as 100μM [24] and even lower with recent techniques such as cryoprobe technology [25]. NMR spectroscopy utilises the fact that under a magnetic field, the atomic nuclei (in this case, ^{1}H) absorb at a frequency proportional to the strength of the field, by detecting the resonance of hydrogen nuclei. When placed in a magnetic field, the magnetic moment of the hydrogen atom adopts one of the two permitted orientations of different energy. The difference in energy of these two states depends on the strength of interaction between the magnetic moment of the nucleus and the field [26]. This energy difference can be measured by applying electromagnetic radiation of a certain frequency which causes the nuclei to shift between states. 'Chemical shift' is defined as the effect of the chemical structure of the molecule on the resonance frequency of the nuclei. Therefore the NMR spectrum for each metabolite is comprised of a characteristic pattern of peaks or resonances, derived from three main factors:

1. The chemical shift (δ) of each resonance is dependent upon the local magnetic field experienced by each nucleus. This local field is dependent on the degree to which molecular orbitals shield the influence of the external spectrometer field. Thus the chemical shift reflects the chemical structure and bonding configuration of the metabolite. Whilst Hz is the fundamental energy unit of NMR, the frequency reported is dependent on the magnetic field strength so the position of each peak is measured relative to that of an internal standard and given in a scale of parts per million (ppm) instead [26]. A commonly used internal standard is 3-(Trimethylsilyl)-Propionic acid-D4 (TSP).

2. Spin-spin coupling (also known as scalar coupling or J-coupling) is the phenomenon of magnetic interactions between nearby nuclei. This means a proton has more than one resonant frequency resulting in a juxtaposition of peaks called a 'multiplet' as shown in Figure 1.1, with the pattern determined by the chemical structure of the molecule.

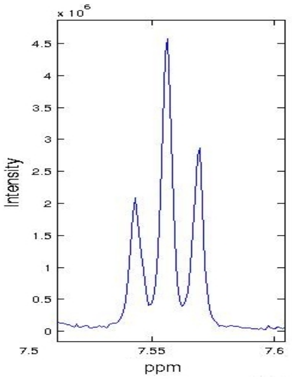


Figure 1.1: Example of a NMR multiplet signal resulting from spin-spin coupling.

3. Integrated peak area for a given metabolite is proportional to the number of observed ^{1}H nuclei (assuming there are no differential relaxation effects) and allows quantification of the metabolite concentrations.

The resulting dataset consists of a series of resonance intensity measurements taken over a grid of frequencies, with the x-axis corresponding to the resonant frequency (usually plotted to increase from right to left) and the y-axis gives the resonance intensity (see Figure 1.2). The spectrum of a pure compound will consist of a 'signature' of peaks reflecting the factors described. Under ideal conditions, each peak takes the form of a Lorentzian curve, represented by the following equation:

$$l_\gamma(x) = \frac{2\gamma}{\pi(4x^2 + \gamma^2)} \tag{1.2}$$

where γ is referred to as the 'linewidth' (or 'peak-width at half-height'). The NMR spectrum of a complex mixture can be well approximated by a linear combination of the spectra of pure compounds, i.e a biofluid spectra containing K different metabolites can be treated as K-dimensional objects, in which each dimension represents the concentration signal of a single metabolite [27]. This superposition of peaks and multiplets results in a complex spectrum of overlapping signature patterns. These spectra can be further complicated by peak positional variation, due to matrix effects or differences in experimental conditions as well as variation in the chemical properties of the sample, e.g pH value and the strength of other ionic species in the

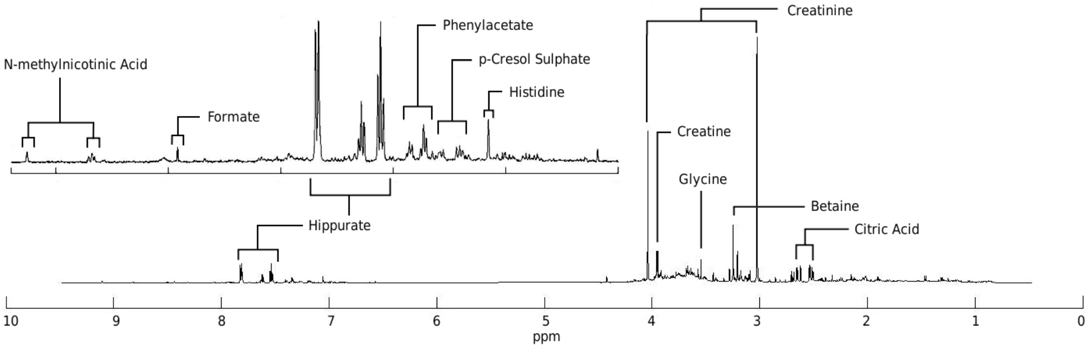


Figure 1.2: A typical ^{1}H NMR spectrum of a human urine sample with some labelled metabolite resonances, where the x-axis is ppm (parts per million) relative to some internal standard and the y-axis gives the signal intensity.

mixture [28]. These problems, along with background noise and the presence of contaminants, make peak identification very difficult particularly for automated algorithms.

Pre-processing is typically performed to provide a dataset that is more informative for statistical modelling. The data is usually transformed into a matrix of m variables (metabolite concentrations or metabolic signals) by n samples. This may be achieved by first identifying individual metabolite signals and quantifying them (both non-trivial tasks), or by 'binning' (also referred to as 'bucketing') the spectra into equal regions (typically 0.04ppm wide). Binning is often preferred to using the entire chemical shift resolution as not only does it reduce the dimension of the data, but may also limit the effect of peak positional variation. Alternatively, variation in peak position can be reduced by using an alignment procedure [29]. Normalisation of the matrix rows may be performed with the aim of removing variation between samples by multiplying by a constant. The samples may be normalised relative to an internal standard or the total integrated intensity across the spectrum, depending on whether it is more important to determine absolute or relative levels of metabolite concentrations. Another option is to normalise the spectra with respect to a selected 'reference' spectra using probabilistic quotient normalisation or histogram normalisation for example, but knowledge of the biological context of the data is essential when choosing a normalisation method [30, 29]. Sometimes the columns of the data matrix will also be scaled, allowing more emphasis to be placed on low intensity signals. Variables are often scaled to unit variance, or alternatively Pareto or Log scaling may be used [25]. Another typical pre-processing step is to remove the spectral region worst affected by the unwanted water signal, since suppression of these large resonances is usually imperfect.

The next step in the statistical analysis of metabolic NMR data is to perform both unsupervised (exploratory) and supervised (e.g. classification/regression) analysis, with the aim of identifying and characterising important differences between classes resulting from the presence of external or internal stimuli to the organism. Classical methods (employed in metabolomic analysis across technologies) include Principle Components Analysis (PCA) and Partial Least Squares (PLS) regression, which aid interpretation of the data by projecting onto lower dimension spaces corresponding to either the highest variance, or highest covariance between the scores of the data and the response variable. Typically the analysis workflow will include firstly unsupervised clustering such as PCA and then be followed by a supervised method (PLS-DA etc.) to enhance this separation and understand which variables are responsible for the differences between datasets.

PCA is a useful exploratory analysis tool. PCA transforms a dataset consisting of observations of several variables into a set of orthogonal 'principal components'. The principal components are ordered in such a way that the first accounts for the as much variance in the data set as possible and the last accounts for the least, whilst maintaining orthogonality between the components. PCA scores correspond to a linear combination of the original variables and can be defined as the transformed variable values corresponding to a particular data point. Loadings describe the weight that each original variable contributes to the component score. Usually, the dimensionality of the problem is reduced by considering only the first two or three principal components as they will typically account for most of the variability of the data. This technique is particularly relevant for metabolomics given the original variables are highly collinear.

PLS is a broad class of techniques for modelling relations between variable sets using latent variables. Generally speaking PLS creates orthogonal score vectors (or components) that maximise covariance between different sets of variables by projecting *both* the predictors and outcomes to a new space. It is particularly useful when the variables outnumber the observations and are highly collinear. Whilst PCA chooses the x-axis scores to explain as much factor variation as possible, PLS chooses x and y scores so that the relationship between successive pairs of scores is as strong as possible as expressed by covariance of scores. As such, PLS scores reflect the covariance structure between predictors and response, with the aim being to maximise covariance between the response variable and component scores. Extensions of these methods, for example PLS regression, Orthogonal (O/O2-PLS) PLS and Discriminant Analysis (PLS-DA), are also used throughout the metabolomic community. Alternative approaches include Support Vector Machines, Genetic Algorithms and Programming, kernel methods and tree-based classifiers (e.g. Random Forests) [30, 25].

Statistical Spectroscopy is a collection of tools that has been rapidly expanding since the development of Statistical Total Correlation Spectroscopy (STOCSY) [31]. STOCSY utilizes correlation analysis to construct a pseudo-2D spectrum providing useful information on correlated metabolic signals in ^{1}H NMR. This is useful for identifying signals belonging to the same compound (some molecules have more than one signal for each measurement) as well as highlighting correlated metabolites which may be part of the same metabolic pathway. The methods used by STOCSY have now been extended to tackle analysis across other technologies, for example Statistical Hetero-Spectroscopy (SHY) aims to incorporate information obtained using MS data [32]. 2D-NMR methods employing an additional NMR nuclei, e.g. ^{13}C, are often used similarly in metabolomic analysis with the advantage of providing further information useful in elucidating the chemical structure and identifying unknown molecules, as well as signal dispersion in the additional dimension [33].

Identification of metabolite signals present in NMR data is greatly aided by the development of several databases dedicated to documenting metabolite data. Two of the largest are the Biological Magnetic Resonance Bank (BMRB) [34] and the Human Metabolome Database (HMDB) [35]. Although the BMRB includes information on molecules including peptides, proteins, and nucleic acids, the HMDB focusses specifically on small molecule metabolites found in the human body. Whilst these databases are valuable tools, there exist several challenges in curating the 'ideal' data resources. Two major problems are the unknown abundance of unknown metabolites within the human metabolome meaning the databases are inherently limited in their coverage, and also the relevance of acquisition conditions, for example there may be much variation in pH, use of solvents etc for different metabolite data. In addition to the metabolite databases for particular technologies there are useful resources linking metabolites to particular pathways, for example Kyoto Encyclopedia of Genes and Genomes (KEGG) [36], MetaCyc [37] and ConsensusPathDB [38].

The methods for processing and interpreting metabolic ^{1}H NMR data are many and varied. In Chapter 2, a package for simulating human urine ^{1}H NMR metabolic profiles is introduced, whilst Chapter 3 tackles the problem of modelling pH-induced peak positional variation.

1.2.2 Liquid Chromatography - Mass Spectrometry

Most biologically interesting chemicals are present as isomers: molecules having exactly the same mass but a different structure. Using a separation technique, such as chromatography, before using mass spectrometry can help tackle this problem by further dispersing the signal in an additional dimension. Another advantage of separation is that it limits the amount of analytes being ionized simultaneously and thus reduces the possibility of ion suppression effects. Ion suppression can also be a result of the sample matrix used and is unpredictable, impairing detection capability, quantification and reproducibility [39, 40]. After elution, the mobile phase passes through an interface before moving to the mass spectrometer. This interface either attempts to remove the mobile phase, or pumps the solution straight into the mass spectrometer in order to minimize contamination. Figure 1.3 illustrates the LC-MS work-flow.

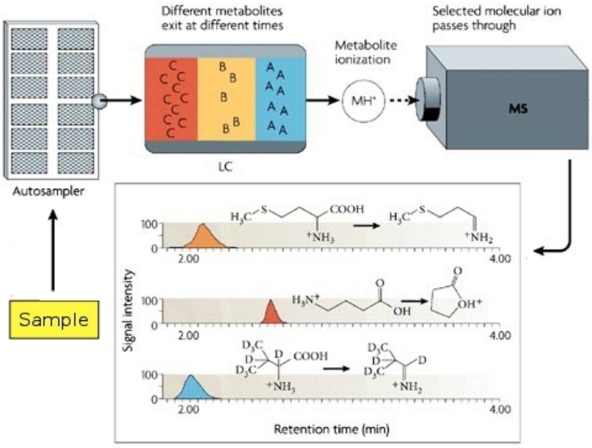


Figure 1.3: A high level view of the LC-MS workflow taken from [1]. The sample enters the liquid chromatography where compounds are separated in the time domain, then the sample is ionized and passed to the mass spectrometer where compounds are further separated by their mass. This results in a 2-dimensional dataset where signals are described by a (retention time (Rt), mass-to-charge ratio, intensity) triplet.

Liquid Chromatography works by pumping a 'mobile phase' solution through the 'stationary phase' or 'column', where the time taken from injection to elution at the detector is dependent on the interaction between the analyte and the mobile and stationary phases. The mobile phase is a solution containing the sample and any solvents or buffers used. Solvents are used in order to control the strength of interaction different molecules have with the chromatographic column, whilst buffers manage the degree of ionization of the analyte. Even high-purity solvents and buffers can contribute contaminant signals to background noise. The mobile phase is then delivered into the column under high pressure in order to sustain a constant flow rate and thus ensure reproducibility. Separation occurs when elements of the sample mixture interact to different extents with either or both of the mobile and stationary phases and so take different amounts of time to traverse the column. There are many different configurations of column and mobile phases that give widely varying results depending on the physiochemical properties of the molecules analysed [41]. Gradient elution, for example, actually varies the composition of the mobile phase during the experiment in an attempt to achieve higher resolution separation.

LC technology is fast evolving and advances in methodology means that the peak capacity of a single scan is ever-increasing. HPLC (High Performance Liquid Chromatography) has been a widely used technique since the 1990s. HPLC utilises the fact that employing smaller particles for the stationary phase of the column and injecting the mobile phase at high pressure increases both the speed of elution and the number of peaks resolved per unit time in gradient separations, resulting in both improved sensitivity and resolution. More recently UPLC (Ultra Performance Liquid Chromatography, pioneered by Waters MS Technologies, Ltd., Manchester, U.K.) has become the fore-running technique using yet smaller particles to improve performance even further. The enhanced resolution and sensitivity means that run times are quicker and reproducibility is better than for normal HPLC.

Mass spectrometers differentiate compounds within a complex mixture by measuring their molecular mass-to-charge ratio. Upon entering the mass spectrometer, the sample is ionized using one of a variety of techniques : electron or chemical ionization, fast-atom bombardment (FAB), thermospray (TSP), electrospray (ESI) etc, with ESI the most commonly used in metabolomics. ESI is a relatively ‘soft’ ionization technique that causes minimal fragmentation and can be run in positve and negative modes in order to capture both sets of analytes. Then, ions of different mass-to-charge ratios (m/z) are separated and the counts of each group of ions is measured. There are several different methods for mass analysis, for instance Quadrupole Mass Analysers vary the voltage between two pairs of rods in order to bring ions of different m/z to the detector as required. This instrument has high sensitivity for targeted analysis [42]. Another type of mass analyser common within metabolomics is Fourier Transform Ion

Cyclotron Resonance (FT-ICR), which determines the m/z value based on the cyclotron frequency (dependent on mass) of the ions trapped in a fixed magnetic field. The signal is Fourier Transformed to the frequency domain in order to calculate the mass spectrum [43]. FT-ICR has a very high mass accuracy so is particularly useful for identifying unknown analytes. Perhaps the simplest mass separation device is the Time-Of-Flight (TOF) analyser. The ions are first accelerated by an electromagnetic field. TOF then utilizes the fact that regardless of the ionization technique, all the ions are given the same kinetic energy meaning that the velocity of each is inversely proportional to the square root of its mass. Thus, the m/z of the ion can be deduced according to how long the ion takes to traverse the flight tube of the instrument [42]. Due to its simplicity and fast scanning capability, TOF is being used increasingly in LC-MS instrumentation to produce high-resolution spectra for untargeted applications.

Mass analysers can also be run in tandem with an alternative mass analyser (or indeed more than one), for example QTOF mass analysers combine the strengths of Quadropole mass analysis with those of Time-Of-Flight analysis simultaneously to improve sensitivity and accuracy. These tandem mass spectrometers can also be used to produce additional structure information about compounds by fragmenting ions and identifying the fragments. The Quadropole analyser may be used to select a mass range which is then measured by the TOF, or alternatively for fragmenting ions which are then detected in the TOF analyser.

The whole process results in a 2-dimensional data set: an intensity for each retention time (Rt) and mass-to-charge ratio (m/z) pair, as shown in Figure 1.4 [42, 44]. Metabolite quantification relies on the fact that peak intensity (or peak volume) is usually proportional to the concentration of a molecule [45]. A ‘mass spectrum’ is defined as the ions detected for one particular retention time value, i.e. a slice in the retention time dimension, whilst slicing at constant m/z gives a ‘chromatogram’ (Extracted Ion Chromatogram or EIC) for each m/z value. In constructing chromatograms, one can either use the maximum intensity (resulting in a ‘Base Peak Chromatogram’) *or* the total intensity at each Rt point in the slice (EIC). However, complications arise in defining a m/z ‘bin’ for each chromatogram since there can be drift in the measurement of m/z (see Figure 1.5) and this problem will be examined more closely in Chapter 4. The characteristics of peaks in the m/z direction or ‘MS peaks’ are largely determined by the

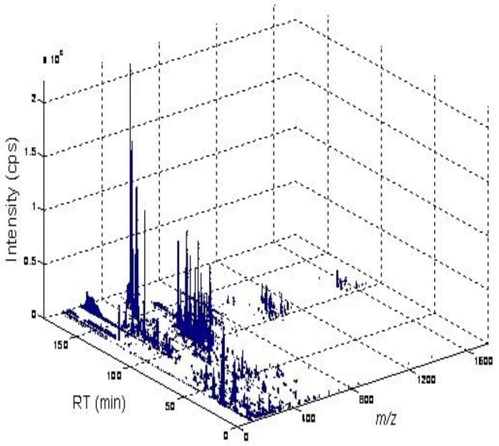


Figure 1.4: A 2-D Intensity map of LC-MS Data from [2]. x and y axes give the Retention Time (Rt) and mass-to-charge ratio (m/z) and z axis gives the intensity (ion count) detected.

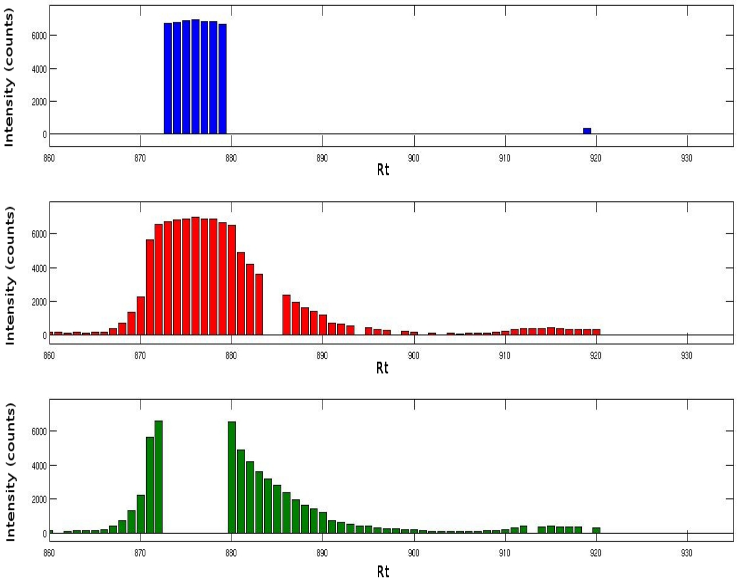


Figure 1.5: A chromatographic peak split across three bins due to drift in m/z.

instrument used. Gaussian curves are commonly used for MS peak modelling, although other shapes may perform better [23]. The LC peak shape is dependent on a complex interaction of factors, not entirely understood, such as solvent concentration, separation gradient etc. Again, many use a Gaussian curve for simplicity [46, 47], but it is acknowledged that peaks usually have a long tail and are sometimes even bi-modal [41, 48]. Chromatographic peaks often require alignment between successive samples due to the difficulty ensuring the reproducibility of the process resulting in non-linear distortions of the Rt axis. The chromatographic domain is usually used in order to discriminate between analyte and noise peaks [45].

The signal is complicated further by isotope patterns (often predictable if the compound identity is known) due to the abundance of naturally occurring isotopes, such as C^{13}. This means multiple peaks may be registered for a single molecular species in mass spectra at different m/z locations. Figure 1.6 shows an example of protein MS. The mono-isotopic peak registers at (54.7s, 526.6m/z), and at least two well-defined smaller isotopic peaks are present at 527.5m/z and 528.3m/z. In addition to this, heavy molecules such as proteins may generate a group of related peaks with different charge states[41]. However, metabolites are relatively small by comparison (<1500Da) producing few, if any multiply charged ions so this is not generally a big problem within metabolomics [42]. Other complications include fragmentation of molecules during the ionisation process, resulting in a parent and correlated daughter ions being detected. Linking fragments (co-eluting peaks) using statistical post-processing techniques can be helpful in identifying unknown molecules. Sometimes dimerization can occur, where two molecules of the same species combine and so are detected as a single ion at twice the m/z value. Ionization can also result in adducts, for example protonated or deprotonated alkali metals binding to-

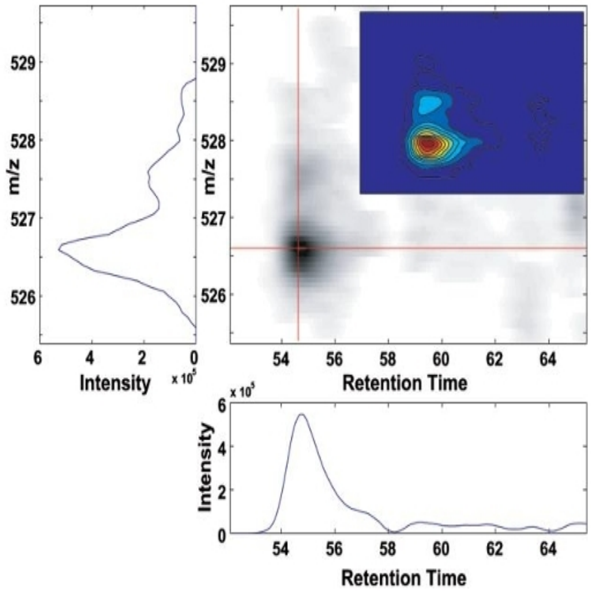


Figure 1.6: An isotopic cluster (proteomic data acquired on a low-resolution spectrometer) as it is visualized in a 2-D map from [3]. The x-axis and y-axis of this noise-filtered 2-D map represent the chromatography and MS dimensions, respectively. Relevant mass spectrum and mass chromatogram are represented as cross-sections. A contour plot of the 2-D map is shown in the inset.

gether as well as neutral losses (water CO_2 etc.) for metabolites. The mechanism behind this process is not fully understood but given the small size of metabolites adduct formation can significantly alter m/z values [49].

There exist numerous obstacles to overcome when interpreting LC-MS data. Tailing peaks can often be mistaken for multiple consecutive peaks, and quantification for one peak can contain contaminating information from a confounding overlap of signals [50]. This can be overcome somewhat by modelling the spectra as a sum of peaks, each represented in some parametric form and then performing deconvolution, but this can introduce new errors. The signal can be corrupted by ionization and ion-suppression effects, usually arising when polar contaminants are competing with the target analyte present for available ions, as well as by impurities from the buffers and solvents used [17].

Noise is usually characterized by high frequency, low intensity, narrow peaks, and is generally more prevalent at the beginning and end of elution. Most molecules elute with an intensity markedly greater than the expected level of noise, but molecules eluting with a low abundance makes them extremely difficult to identify. Feature detection is hampered further by non-linear drifts in retention time (and sometimes m/z values as mentioned earlier). In addition, aligning spectra across many samples is complicated by peaks that are missing or simply undetected in some samples, providing another huge challenge beyond the scope of this thesis [45].

Similarly to NMR there exist reference databases for MS spectra, such as MassBank [51] and METLIN [52]. However, since the between-sample variation of elution times for even the same instrument can be so large, no Rt index is currently available for LC-MS [53]. In adition to spectral library matching, a combination of techniques, such as NMR, are usually required to gain enough structural information to confirm a metabolite identity. Sample preparation method, choice and configuration of LC-MS instrumentats and algorithm/software employed for data processing can result in widely varying datasets even from the same sample. Pharmaceutical regulators are moving to introduce standard protocols in an effort to make results more standardised and comparable (Metabolomics Standard Initiative) and use of pilot studies, internal standards, QCs and other 'quality assurance' measures are widely used in order to tackle the issue of reproducibility and robustness of results.

1.3 Bayesian Inference

Both NMR and LC-MS require sophisticated statistical techniques in order to perform inference on the data obtained. Within statistics there are two main inferential frameworks: ‘frequentist’ and ‘Bayesian’. Frequentist inference is based upon the assumption that the observed data can be considered as one instance of a series of infinitely repeatable experiments [54] and standard frequentist methodologies include statistical hypothesis testing and calculating confidence intervals. Whilst frequentists tend to consider the probability of the observed data arising given a hypothesis, Bayesians are more converned with the probability of a hypothesis given the observed data. Bayesian inference is derived from Bayes’ Theorem (Equation 1.3):

$$p(\theta|y,\phi) = \frac{p(y|\theta)p(\theta|\phi)}{p(y|\phi)} \tag{1.3}$$

where $p(y|\theta)$ is the likelihood of the data given the model parameters, θ is a parameter of the likelihood distribution with prior $\theta \sim p(\theta|\phi)$, ϕ is a hyper-parameter of the distribution of θ, $p(y|\phi)$ is the marginal likelihood:

$$p(y|\phi) = \int_\theta p(y|\theta)p(\theta|\phi)d\theta \tag{1.4}$$

and $p(\theta|y,\phi)$ is the posterior probability of parameter θ.

Practically speaking this means setting up a full probability model, conditioning on the observed data (y) in order to calculate the posterior distribution (the conditional probability distribution of the unobserved quantities of interest, given the observed data) and then evaluating the model fit [55]. Frequentist ‘results’ are usually true or false conclusions drawn from significance tests, whereas Bayesian results more often take the form of probability distributions for parameters that attempt to describe the data. Bayesian methodology is employed widely within metabolomics for a variety of different purposes, for example variable selection/dimension reduction, latent variables analysis, network/pathway analysis and spectral deconvolution among many others.

When practicing Bayesian inference, it is often useful to be able to calculate posterior estimates of characteristics of the model parameters. In the case of multi-parameter models, where $\theta = (\theta_1, ..., \theta_k)$, this requires averaging over 'nuisance' parameters (parameters on which one is not concerned with performing inference). Supposing the parameter of interest is θ_1, the conditional distribution $p(\theta_1|y)$ must be derived from the joint posterior distribution $p(\theta|y) = p(\theta_1, ..., \theta_k|y)$. Averaging over the nuisance parameters gives:

$$p(\theta_1|y) = \int_{\theta_k} ... \int_{\theta_2} p(\theta_1, ..., \theta_k|y) d\theta_2 ... d\theta_k \tag{1.5}$$

or alternatively

$$p(\theta_1|y) = \int_{\theta_k} ... \int_{\theta_2} p(\theta_1|y, \theta_2, ..., \theta_k) p(\theta_2, ..., \theta_k|y) d\theta_2 ... d\theta_k \tag{1.6}$$

However, these posterior distributions are often high dimensional and very difficult to calculate either analytically or numerically. The problem of making inferences on this type of distribution is addressed by using Markov Chain Monte Carlo (MCMC) methods. MCMC enables simulation of random draws from a complex probability distribution, say $f(x)$. MCMC methods are based on Markov Chains, which are stochastic processes X_t, $t = 0, 1, 2, ...$ such that:

$$P(X_n = x_n|X_0 = x_0, ..., X_{n-1} = x_{n-1}) = P(X_n = x_n|X_{n-1} = x_{n-1}) \tag{1.7}$$

i.e. the current observation only depends on the previous one and not the entire observation history [54]. MCMC methods essentially involve constructing a Markov Chain whose target distribution is the distribution of interest $f(x)$. Once a large enough sample has been simulated, the functionals of interest can be calculated to any degree of accuracy.

There are numerous methods for building the required Markov Chain. For example Metropolis-Hastings uses a 'proposal distribution' $q(x)$ to propose a candidate value Y for X_{t+1}, possibly depending on X_t, and accepts it with probability $\alpha(X_t, Y)$ where

$$\alpha(X_t, Y) = min(1, \frac{f(Y)q(X_t|Y)}{f(X_t)q(Y|X_t)}) \qquad (1.8)$$

and $f(x)$ is the function of interest. A special case of the Metropolis-Hastings is the Gibbs sampler. Suppose draws of $\theta = (\theta_1, ..., \theta_k)$ must be obtained from the joint distribution function $f(\theta_1, ...\theta_k)$. Then for each draw $t = 1, 2, ...$, each $\theta_i^{(t)}$ is sampled from the conditional distribution given by $p(\theta_i^{(t)}|\theta_1^{(t)}, ..., \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, ..., \theta_k^{(t-1)})$ (proportional to the joint distribution) so that each variable is sampled using the most recently updated values of the other variables. There are many variations on these samplers, so-called 'adaptive' modifications aimed to increase the sampling efficiency of the algorithms.

A 'burn-in' period is usually utilised (i.e. the first M iterations are discarded) as we accept we are unlikely to choose good starting conditions and thus the initial estimates are likely to be poor. Thinning is the practice of saving only every n-th iteration sometimes with the aim of speeding up post-processing or reducing required memory but also as an attempt to remove auto-correlation. For example, to obtain a run of 10,000 iterations one would run $n \times 10,000$ simulations and save only every n-th one. However it has been argued that if the entire chain is long enough, the auto-correlation has likely averaged out anyway and thinning provides little additional benefit to this end [56].

Convergence of MCMC is a major concern for Bayesian statisticians with the parameters estimated often correlated with themselves over the iterations or with each other making convergence slow and difficult to execute. Despite much theoretical research into convergence computations, there is limited benefit thus far for practical applications. Although it remains impossible to be certain that your MCMC sample is truly representative of the target distribution, there are many diagnostic measures and techniques to help somewhat evaluate success. Indeed Cowles and Carlin provide thorough reviews of these[57].

One of the simplest checks is visualisation of how well your chains are 'mixing' or moving around the sample space. Plotting traces (parameter value vs. iteration number) can show you if your parameter gets stuck or moves poorly. ‘Running mean plots’ (plotting the sample mean up to each iteration vs. iteration number) can also be useful to this end. Investigating the auto-correlation can also provide clues as to the convergence of your chain. A ‘k-th lag auto-correlation’ can be computed and we would expect the correlation to decrease as the lag increases. Persistently high auto-correlation for high k is again indicative of poor mixing.

One commonly used convergence diagnostic is the ‘Gelman and Rubin Multiple Sequence Diagnostic’ which is calculated using multiple chains per parameter. Consider the ‘potential scale reduction factor’:

$$\hat{R} = \sqrt{\frac{V\hat{a}r(\theta)}{W}} \tag{1.9}$$

where $V\hat{a}r(\theta)$ is the estimated variance of the target distribution as a weighted average of the within-chain, W, and between-chain, B, variances:

$$V\hat{a}r(\theta) = (1 - \frac{1}{n})W + \frac{1}{n}B \tag{1.10}$$

When $\hat{R}$ is high (> 1.1 or so) this indicates the chains should be run longer to improve convergence. A ‘Gelman plot’ shows how the potential scale reduction factor changes through the iterations and is another useful diagnostic.

There are several software packages available for automating the Bayesian analysis of models via MCMC methods. WinBUGS (arising from the BUGS - ‘Bayesian inference Using Gibbs Sampling’ project based in the MRC Biostatistics Unit at Cambridge, England) is one of the most popular [58]. It provides a framework for defining Bayesian hierarchical models and a library of sampling routines to perform inference. Several extensions have been developed allowing construction and analysis of ever more complicated models. One such alternative, JAGS (Just Another Gibbs Sampler), was developed with the objective of being an open-source engine for the BUGS language [59]. JAGS is highly extensible and allows users to develop new libraries and add-ons. Both WinBUGS and JAGS can interface to R, making them useful tools in the arsenal of any Bayesian statistician. Whilst MATLAB was used extensively during the development of this thesis to write new algorithms, both WinBUGS and JAGS were used (via R) to perform Bayesian statistical analyses.

Bayesian methods are used throughout 'omics' analysis from processing and analysing NMR [60], LC-MS [61, 62] and GC-MS [63] data among others, to elucidating metabolic pathways [64]. Within metabolomics, particularly when modelling data that is driven by some complex physio-chemical reaction, the Bayesian framework allows us to incorporate prior knowledge of this underlying mechanism whilst allowing the observed data to speak to us about the other unknown interactions occurring that result in both structured and unstructured variability.

1.4 Aims

The major bottlenecks within metabolomics research can be broadly categorised as metabolite identification, data processing and interpretation of results. Metabolite identification depends on the robustness of the data capture, ability to match with standards and is complicated by the great variability of molecular structures and abundance. Data processing and reduction techniques are complex and often bespoke depending on the focus area of a given laboratory, meaning that different results can be produced from a single dataset by using alternative software or statistical methods hampering reproducibility and validation by different research groups. In addition to this, manufacturers' processing software is often tied to their particular spectrometer, although open source software developers and researchers attempt to produce algorithms that perform competitively whilst being compatible with multiple machine types used across different labs. Further downstream in the data flow, all these issues greatly affect the ability to assign biological significance to experimental results, i.e. the ultimate aim of most metabolomic investigations. Given the complex nature of the numerous statistical challenges facing metabolomics research, this thesis sets out to tackle a few of these problems.

- In order to fulfil the need within metabolomics for simulated ^{1}H-NMR data with which to develop and test new techniques for statistical inference, Chapter 2 presents a novel package 'MetAssimulo', designed to draw on data from the Human Metabolome Database (HMDB) and an NMR Standard Spectra Database (NSSD) in order to simulate realistic human urine spectra [65]. It also allows inclusion of peak shift and inter-metabolite correlations.

- Chapter 3 aims to further demystify ^{1}H-NMR metabolic profile data by Bayesian modelling of the relationship between pH and peak positional variation. Being able to estimate peak shift parameters would be of great use in the deconvolution of NMR spectra and also accurately simulating peak shift using MetAssimulo.

- In Chapter 4 I move on to the challenges faced in the processing of LC-MS data, taking a look at existing algorithms and the variety of techniques used. A Bayesian Change Point Model is also explored for this application. H-MS, a novel method for data-binning and peak detection, is described and its performance is compared to an adaptive technique apLC-MS [66] and a commonly used package XCMS [46].

Chapter 2

Simulation of Realistic ^{1}H-NMR Metabolic Profiles: MetAssimulo

Metabolic NMR spectra are highly complex and the field benefits greatly from the application of machine learning and other statistical tools to extract information. Pattern recognition analyses such as Principal Components Analysis (PCA) have long been combined with NMR to investigate normal and pathological metabolic states [67]. Data processing methods are being developed to extract metabolite information and concentrations from raw spectra, allowing automation of spectral processing. Development of advanced mathematical, statistical and computational methods are also essential for characterisation of the metabolic state, delineation of metabolic changes over time and the efficient identification of potential biomarkers. There are a wide variety of diseases where key changes in metabolites have been deduced e.g. cancer, diabetes, hypertension etc. [68, 69, 14]. However, as algorithms and methods are developed, they need to be refined and validated to ensure results will be biologically meaningful. It is hard to effect this without using test datasets where the true answers are known; this can be accomplished using simulation techniques. An alternative approach is to design artificial mixtures of metabolites which are prepared and analysed in identical fashion to real samples. However this is expensive in terms of man power and instrument time, and offers few advantages over in silico simulation when assessment of analytical procedures is not required.

The purpose of MetAssimulo is to simulate datasets of realistic NMR spectra with *known* parameters in order to test data analysis techniques, hypotheses and experimental designs. Few methods for generating simulated NMR datasets have appeared in the literature to date [70, 71]. Most model a limited number of metabolites, make no attempt to reproduce realistic levels of metabolites, and do not allow for between-metabolite or 'inter-metabolite' correlations ([71] excepted) and do not always model peak positional shifts. It is common to fit Lorentzian peak shapes in an attempt to characterize spectral peaks, e.g [70]. However, this ignores the fact that peak shapes in real NMR profiles are variable and can be far from ideal. Here a novel approach is outlined: making use of individual standard metabolite data extracted from the Human Metabolome Database (HMDB) [35] and a local NMR standard spectra database (NSSD). Many metabolic profiling labs host their own NSSD appropriate to the biological systems and sample types they work with and thus the simulations can be tailored to virtually any sample type or organism as required. In this work human urine is used as the example biofluid as it is one of the most widely used in the field and, in healthy subjects, has no protein or lipid content, both of which make the simulation more complex. MetAssimulo is written in MATLAB with a graphical interface allowing the user to alter processing parameters and add new standard spectra as needed. The software is freely available along with an example NSSD of 48 metabolites commonly found in normal human urine (at `http://cisbic.bioinformatics.ic.ac.uk/metassimulo/`). It must be stressed that this list of metabolites and their concentration means and standard deviations does not constitute a *definitive* description of human urine; such a goal is beyond the scope of this chapter. It is provided for the sole purpose of demonstrating the capabilities of the software.

MetAssimulo was initially developed by a student Rebecca Jones. Rebecca worked on the first version of MetAssimulo able to simulate peak shifts. The author's work built on this: bug fixes, incorporating metabolite correlations, extracting data from the HMDB and increasing efficiency which required a substantial amount of recoding. This chapter is based on the MetAssimulo article published in BMC Bioinformatics [65].

2.1 Implementation

MetAssimulo performs various functions accessed though the Graphical User Interface (GUI): pre-processing the pure spectra, simulating metabolite concentrations, incorporating peak shifts and creating the final mixture spectrum, shown in Figure 2.1. By default it produces two groups of spectra based on different metabolite mixtures; these could represent controls (normal) and cases (diseased) subjects. Each metabolite has a characteristic pattern of peaks on a linear scale, the chemical shift, given by δ in ppm. The signal intensity, $y(\delta)$, in a spectrum of metabolites $k = 1, .., K$ (where K is the total number of metabolites in the mixture) at a given δ increases proportionally to the concentration of each metabolite, c_k, present in the sample and their number of observed protons, p_k. The different metabolite spectra are summed together to produce the overall mixture spectrum. Normally distributed additive noise $\epsilon(\delta) \sim N(0, \sigma^2)$ (see

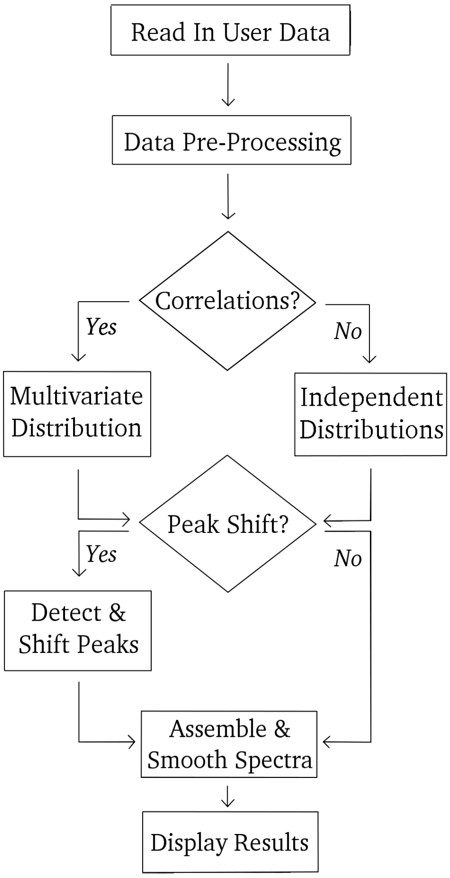


Figure 2.1: A high level flow chart of the MetAssimulo algorithm.

'Calculate noise standard deviation' section for estimate of σ^2) is then added to the mixture spectrum $y(\delta)$, as in Equation 2.1:

$$y(\delta) = \sum_{k=1}^{K}(y_k(\delta)c_k p_k) + \epsilon(\delta) \tag{2.1}$$

This is is then smoothed to simulate the conventional preprocessing technique of exponential apodization prior to Fourier Transform [72]. Each individual metabolite spectrum is sampled at a series of n uniformly spaced data points. The overall spectrum is made up of pairs of data points, $(x, y) = (x_i, y_i)_{i=1,...,n}$, where $y_i = y(x_i)$; n is set by the user, and x_i defines points sampled from the chemical shift δ.

In real NMR spectra, the signal intensity is affected by the extent to which the observed nuclei are allowed to relax before each observation. In MetAssimulo I do not currently attempt to simulate the effects of differential inter-molecular relaxation. However, intra-molecular relaxation effects are accounted for by the fact that experimentally obtained pure compound spectra are used to form the mixture spectra.

2.1.1 Setting Parameters

Parameters can be altered either in the MetAssimulo GUI or within the parameter file 'parameters.txt'. The interface provides the user with several different processing options. For example the second group may be specified as fold change ratios of the concentrations of the first group and the user can specify whether to produce output with or without peak shifts or both. The user also chooses whether to include inter-metabolite correlation (pairwise correlations between metabolites) or not; either as a textfile whose entries can be altered using the interface or constructed from scratch in the Correlation GUI.

Input Files

The Human Metabolome Database (HMDB) [35] contains information about more than 2180 metabolites found in humans and includes literature data relating to normal and abnormal concentrations in biofluids. ‘Metabocards’ is the flat file download of the entire database, available at www.hmdb.ca. Also required is the HMDB set of NMR Peak Lists (containing locations of individual peaks for metabolites) which is available in a downloadable zip-file. In constructing the template of normal human urine concentrations various problems of incompleteness and/or ambiguity were encountered. For example, in many HMDB entries the metabolite concentration mean and standard deviation is unavailable, or simply a range is given. In these cases the standard deviation was estimated by dividing the mean (or ‘half-range’) by 1.95. There are instances where a metabolite is identified as present in urine, but a normal concentration value is not available. As many of these discrepancies as possible have been manually rectified in the provided concentration file by cross-checking with other sources, i.e. literature articles, but do not claim the result represents a complete description of human urine; it does, however serve to demonstrate the software.

The quality of MetAssimulo simulations is also dependent on the quality and coverage of the NSSD used, as well as the peak shift settings affecting multiplet detection. By distributing an NSSD it is not our intention to provide a comprehensive NMR standard database but merely an initial set of common metabolite spectra with which users can begin to make their own simulations. Many users will wish to add their own locally acquired standard spectra for metabolites specific to their areas of interest and I have provided functionality to do this. There are a number of input files that are required for MetAssimulo.

- **Concentration files*** are needed for both groups of metabolites, these detail the mean and standard deviation of the concentration for each metabolite.

- **An NMR Standard Spectral Database (NSSD)** comprising standard 1D ^{1} H-NMR spectra for metabolites is essential. MetAssimulo is designed to work with any metabolite database set out in the Bruker file format. Standard spectra of 48 of the most abundant metabolites in normal human urine is distributed with MetAssimulo.

- **Experiment file** identifying the experiments to use in the metabolite database, as one metabolite may have many spectra, taken at different pH for example.

- **Proton file** listing the number of protons, p_k observed for each metabolite, k.

- **Multiplet data files*** specifying the position of each peak in a multiplet for each metabolite in order to incorporate simulated peak shifts. Known pK$_a$ values and acid/base limits can also be included.

- **Inter-metabolite correlations** can be input via a text file or the GUI.

- **Synonym files*** that allow MetAssimulo to match metabolites in the HMDB data to those in the NSSD.

- **Parameter file** containing the default values or simulation parameters (alterable in the GUI).

* Denotes files which can be generated automatically using ‘Format HMDB Data’ function within MetAssimulo.

Examples of all input files in the appropriate format are included with the MetAssimulo distribution. Much of the required input data can be generated using the in-built function ‘Format HMDB Data’ (accessed via the GUI) which should be run as an initial ‘setup’. It produces the files necessary for conversion between the local database and HMDB synonyms, data required for peak shift simulation and a raw template of concentration data for‘normal’ urine. The normal urine concentration file provided with the distribution has been hand curated to provide realistic values and correct a number of errors found in the current version of the HMDB whilst reducing the number of metabolites used in order to decrease processing time.

2.1.2 Pre-processing

Initially, a set of metabolite concentrations is simulated for the case and control groups, based on the mean and standard deviations in the concentration file. Next, the required spectra from the NSSD must be loaded. Even ^{1}H-NMR spectra of standard pure compounds contain a number of complexities, such as chemical and electronic noise, phase and baseline errors, contaminants and water suppression residuals. Thus it is ususally necessary to preprocess these spectra into a form suitable for combining into the final metabolic profiles.

Simulating Concentrations

Concentrations, c_k, for each metabolite, $k = 1, .., K$, are simulated for the number of replicates specified by the user. Individual metabolite concentrations are generated from a truncated normal distribution, Equation 2.2, using the inverse cdf method since negative concentrations are unphysical [73]:

$$c_k \sim N(\mu_k, \sigma_k^2) I(c_k > 0) \tag{2.2}$$

where μ_k is the mean concentration and σ_k is the standard deviation input by the user for metabolite k.

Significant inter-metabolite correlations, here assumed to be linear pairwise correlations between metabolites, are often found within the field of metabolic profiling so they were considered an important feature to incorporate into the simulation. Where inter-metabolite correlations are required, the concentrations are simulated by sampling from the appropriate multivariate normal distribution. Using the method detailed in [74] the nearest positive semidefinite correlation matrix is calculated given user-specified pairwise correlations. The covariance matrix is constructed using the metabolite standard deviations and specified correlations, and the diagonal entries are increased sufficiently to ensure positive-definiteness. Any necessary alterations to the correlation and covariance matrices are output for inspection.

Read in spectrum

After the concentrations have been simulated the standard spectra of the metabolites are read in. Each spectrum consists of chemical shift in ppm, x and intensity, y. Spectra are then linearly interpolated onto a ppm grid of user-specified resolution.

Exclusion regions

Exclusion regions, corresponding to the location of the internal standard peak (default $<$ 0.2ppm [75]) and the residual water peak (default 4.5ppm - 6.0ppm [75]), are set to zero. In urine, the urea signal (between 5.4ppm and 6.0ppm [75]), the most abundant proton-containing metabolite [27], can be problematic particularly when water-suppression methods are used. Water-suppression is usually imperfect and the resulting residual peaks (near to the urea signal) are not dealt with easily by baseline correction algorithms [30]. Often, the urea and water peaks are combined into one exclusion region lying between 4.5ppm and 6ppm (default exclusion region, but can be adjusted by the user). Excluding these areas of the spectrum helps reduce sensitivity to artifacts.

Baseline Correction

It is easier to distinguish peaks in a spectrum when the baseline is featureless [72], however, spectra can have distorted baselines due to imperfections in the detection process [75]. Curved baselines can be a major source of error and so a correction is carried out on the raw spectrum using a moving average [76]. This method involves splitting the data into windows of size ω (default is 0.3125ppm), defined by the user, then using the median within the window to estimate the baseline. In order to alter the baseline without losing metabolite peaks, a threshold is set by dividing the maximum height by a user specified parameter (default is 10). All the intensities found below this threshold are corrected by subtracting the estimated baseline.

Removal of Negative Artifacts

Negative artifacts, produced by baseline correction or simply inherent in the original spectrum must be removed since their presence could interfere with peaks of interest in the mixture spectrum. This is remedied by using an estimate of the noise standard deviation, σ_{med}, to calculate a limit value, l, using Equation 2.3:

$$l = M - 3\sigma_{med} \tag{2.3}$$

σ_{med} is estimated by splitting the spectrum into a number of bins (given by the user, default 32) and calculating their standard deviations. The median of these standard deviations is used as the estimate of σ_{med}. M is the median of the intensities, y_i. All intensities appearing below this limit,l, are set equal to it.

Kernel Smoothing

Noise from each standard metabolite spectrum will remain in the final mixture spectrum, reducing the overall signal to noise ratio. Kernel smoothing is used to reduce the noise in each individual metabolite spectrum. This process estimates the smooth function underlying the noisy data using a weighted mean of surrounding data points with weights defined according to the choice of kernel. Whilst the default kernel type is 'Normal', the user may also choose from a number of options and also alter the bandwidth (given as number of data points), controlling the degree of smoothing required. Since smoothing the whole spectrum would increase the peak widths, only intensities below a user-defined threshold (a percentage of the maximum intensity, default 0.8 %) are subject to kernel smoothing.

Peak Shift

If the user has chosen to simulate peak shifts (this is the default setting), these are then calculated for each multiplet in each metabolite spectrum. First, a peak detection process is used to identify peaks suitable for shifting. Peaks detected are cross-referenced with the HMDB multiplet data to determine those belonging to a multiplet that must be shifted together. Whether or not a peak is shifted depends on the user defined thresholds for peak detection, and also its size relative to the noise. In real samples, peak positional variation can derive from various matrix effects primarily pH differences but also due to variation in the concentration of other ionic species in the mixture. In MetAssimulo, pH variation only is accounted for; this is sufficient to produce very realistic shift patterns and avoids the need for many extra parameters in the model. If pK_a values and acid and base limits are not available, values are drawn from normal distributions with user-specified mean and standard deviation. If the user requires the same pH for all replicates of a mixture, the pH value is set as the user input. Otherwise, the pH values are sampled from a normal distribution with mean and standard deviation defined by the user. This information is then combined using the Henderson-Hasselbalch Equation [26] (Equation 2.4) to calculate the peak shift (in ppm) and the peaks are shifted accordingly:

$$\eta_{ij} = \frac{10^{pH-pKa_j}(a_{ij} - \delta_{ij}) + (\delta_{ij} - b_{ij})}{10^{pH-pKa_j} + 1} \tag{2.4}$$

where δ_{ij} is the un-shifted position of peak i of metabolite j in ppm (known), η_{ij} is the amount the peak is shifted in ppm (generated by Eq.2.4), pH is the pH of sample (simulated or input), pKa_j is the pK$_a$ of metabolite j (simulated or input) assumed here to be the same for all peaks of a given metabolite, a_{ij} is the position of peak i of metabolite j in the acid limit (ppm) (simulated or input), b_{ij} is the position of peak i of metabolite j in the basic limit (ppm) (simulated or input). See Chapter 3 for more details on peak shift behaviour.

After this process, the spectrum is then smoothed again in order to suppress any unwanted artifacts created by the peak shift.

2.1.3 Simulating Mixture Spectra

To make sure that all the metabolite spectra are on a comparable scale the spectra are normalised to unit integrated intensity, using Equation 2.5:

$$y_i = \frac{\tilde{y}_i}{\sum_{i=1}^{n} \tilde{y}_i} \tag{2.5}$$

where $\tilde{y}_i$ is the intensity in the preprocessed, unnormalised standard spectrum.

Calculate noise standard deviation

The final mixture spectrum is constructed using Equation 2.1. The standard deviation σ of noise to be added is calculated by dividing the maximum peak intensity by the signal to noise ratio required by the user, SNR, as in Equation 2.6:

$$\sigma = \frac{max(y(\delta))}{SNR} \tag{2.6}$$

Even after preprocessing, the signal to noise levels of the individual metabolite spectra may vary, so the final signal to noise ratio cannot be controlled perfectly. However, adding noise in this way allows the simulation of mixture spectra with a wide variety of signal to noise ratios. After adding the random noise, $\epsilon(\delta) \sim N(0, \sigma^2)$, kernel smoothing is used on the composite spectrum to reproduce the effect of apodization on real spectra [72].

2.2 Results and Discussion

In this section some example outputs from MetAssimulo will be shown. Simulations of normal urine were run using the optimized template with parameters set to the default values and using the NSSD consisting of 48 spectra recorded at 600MHz ^{1}H observation frequency.

Single Spectrum

To test whether MetAssimulo's output spectra (Figure 2.2(a)) seem realistic they are compared to a real normal human urine spectrum with the same exclusion regions (Figure 2.2(b)). It should be noted that differences between real and simulated spectra will result not only from the simulation process, but also from incomplete knowledge of the exact molecular species giving rise to NMR signals and uncertainties in their levels. However, despite these difficulties, the simulated and real spectra show many similarities including the dominance of high abundance metabolites such as creatinine, glycine, and citric acid. The insets show how such realistic simulation extends to low intensity signals of the aromatic region such as hippurate, histidine, formate and N-methylnicotinic acid.

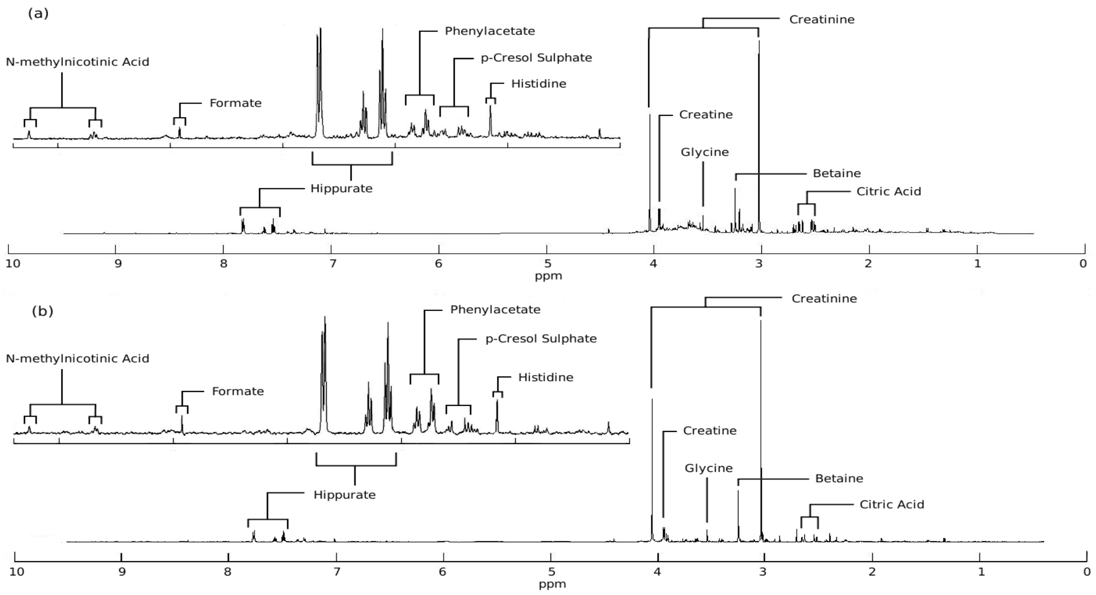


Figure 2.2: (a) Real normal urine ^{1}H-NMR spectrum, (b) Mean of simulated normal urine ^{1}H-NMR spectra produced using MetAssimulo.

Simulation of Case & Control Groups

MetAssimulo can produce two groups of spectra with different metabolite compositions, illustrated here by simulation of spectra for both normal urine and a diseased state. Paraquat poisoning [77] was chosen as the diseased state from several available in the HMDB, because it shows a diverse array of metabolic disregulation in comparison with normal urine. The concentrations of 4 metabolites are altered: citric acid and creatinine are decreased, whilst alanine and lactic acid are increased. Simulations were run for 50 replicates of normal and diseased without peak shifts, the mean of which are shown in Figure 2.3. These spectra clearly show the expected decrease in citric acid and creatinine concentrations for Paraquat poisoning, whilst alanine and lactate concentrations are increased. The PCA scores plot in Figure 2.4(a) clearly demonstrates separation in the first principal component. The first and second principal components (PC1 and PC2) explain 87.5% and 7.0565% of the dataset variance. The largest loadings on PC1, Figure 2.4(b), correspond to the metabolites that were altered, accurately describing the difference between the two groups. The largest loading on PC2 corresponds to glycine, the metabolite with the highest within-group variance. This data could be used in disease diagnostics to help train machine learning methods in recognising disease status.

Peak Shifts

Peak shift is demonstrated using histidine, a metabolite particularly prone to this kind of positional variation. Acid and base limits were estimated by inspecting spectra taken at varying pH values. Figure 2.5 clearly shows a shift in ppm values for this peak consistent with the non-linear mechanism described by the Henderson-Hasselbach Equation 2.4.

Inter-Metabolite Correlations

To demonstrate the incorporation of inter-metabolite correlations, the pairwise Pearson correlations of three metabolites: citrate, creatinine and 2-oxoglutarate are specified. Figure 2.6 shows the correlation matrix used.

This resulted in a positive definite covariance matrix, so no adjustments were required. Figure 2.7(i) visualises the correlation matrix between all spectral intensities. Most correlations are close to zero as expected. The regions enlarged in (ii)-(v) illustrate the the correlations that were expected. The correlations can also be viewed in Figure 2.8 when the mean spectrum is coloured according to the correlation coefficient with respect to a specified chemical shift corresponding to a particular metabolite peak position ((a) citrate-2.65ppm, (b) creatinine-4.08ppm, (c) 2-oxoglutarate-2.44ppm). Note that these analyses are similar to the commonly used STOCSY [31] technique which is used to analyse both inter- and intra-metabolite correlations; our simulations could be used to develop and test such methods.

2.3 Conclusion

There are currently simulation programs in different areas of post-genomic science, such as SNP simulators that are being used in whole genome association studies [78, 79]. MetAssimulo is a valuable addition to these tools, enabling the simulation of realistic ^{1}H NMR spectra of complex biological mixtures including group-wise variation, intermetabolite correlations and peak positional variation. However, there are areas which could be enhanced. Any simulator of this kind is limited by the sources of data available. The HMDB only contains information about metabolite concentrations in humans, therefore further user input or other metabolite databases may be needed to address other organisms. Human urine is the default setting for MetAssimulo, but given the numerous alterable parameters, it is easy to simulate profiles for other species and biofluids.

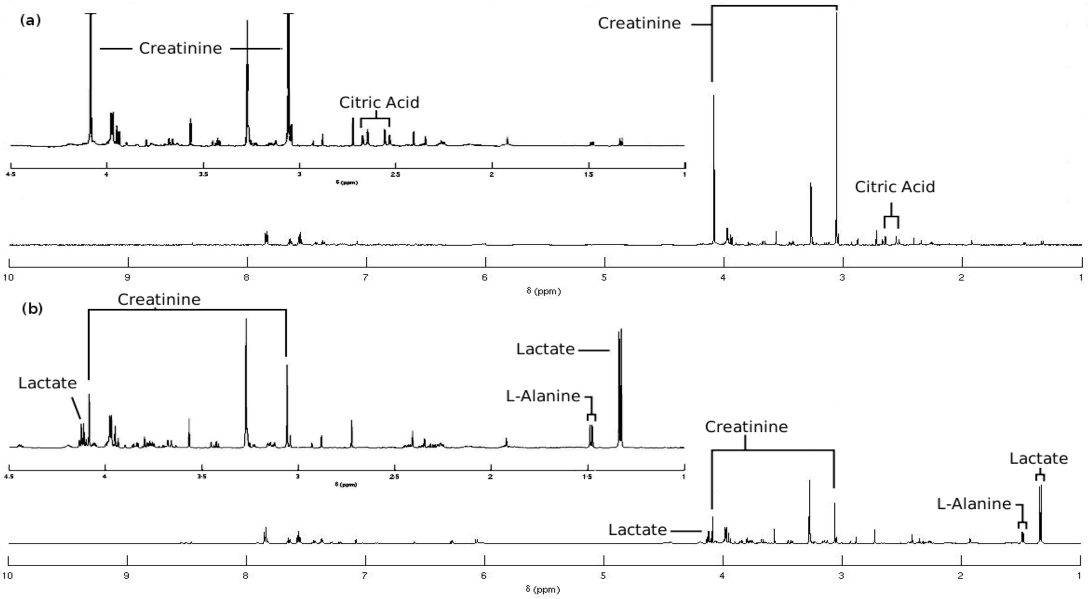


Figure 2.3: (a) Mean simulated normal urine ^{1}H-NMR spectrum, (b) Mean simulated ^{1}H-NMR spectrum of urine in paraquat poisoning produced using MetAssimulo.

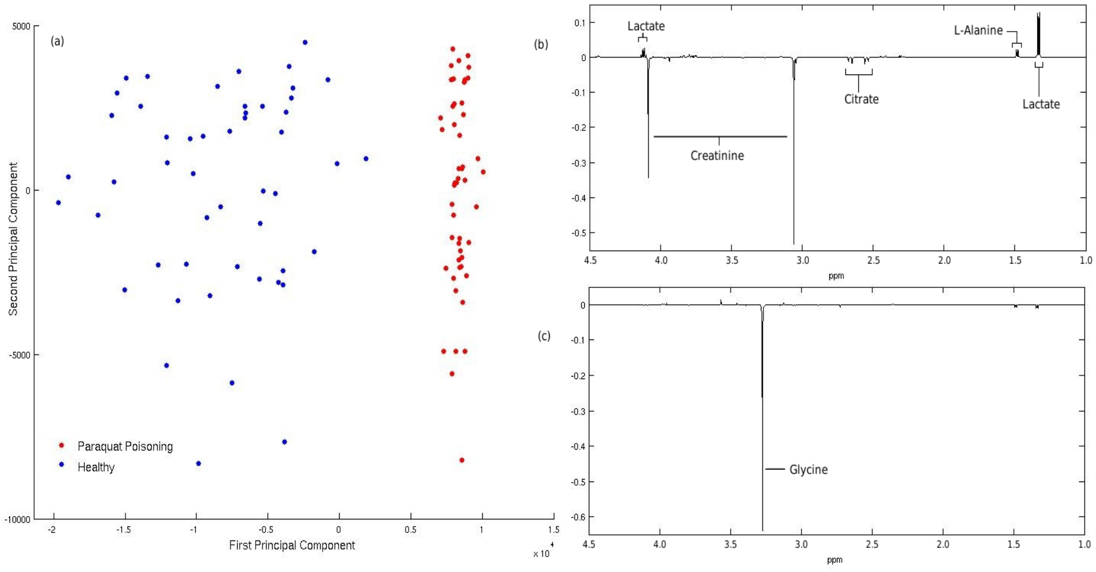


Figure 2.4: (a) PCA scores plot of the first two principal components for the simulated normal and diseased datasets show clear separation, (b) Loadings on PC1 indicating metabolite resonances that describe a large portion of the difference between the normal and diseased datasets, (c) Loadings on PC2 indicating metabolite that has a large within-group variance.

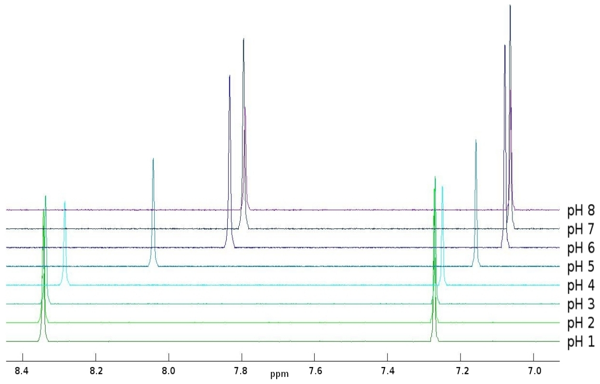
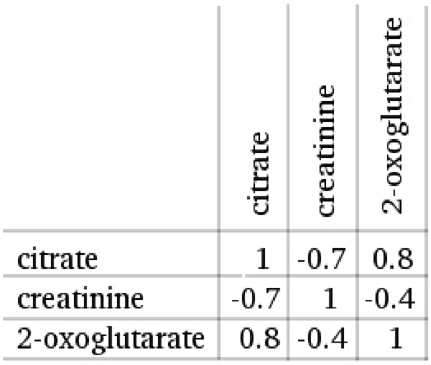


Figure 2.5: Simulated ^{1}H-NMR spectral peak shift for the two aromatic singlets of histidine.



	citrate	creatinine	2-oxoglutarate
citrate	1	-0.7	0.8
creatinine	-0.7	1	-0.4
2-oxoglutarate	0.8	-0.4	1

Figure 2.6: Correlation matrix used to demonstrate MetAssimulo correlation functionality.

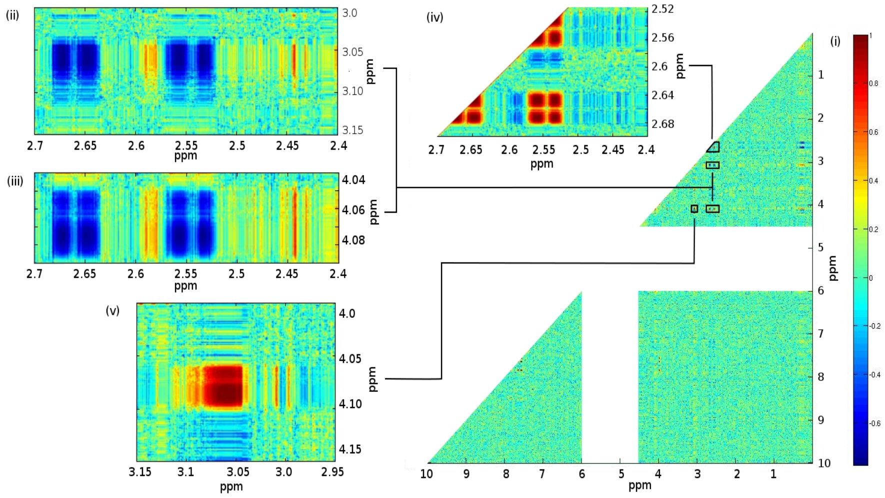


Figure 2.7: Inter and intra-metabolite correlations: (i) Complete correlation matrix and insets (ii)-(v) showing strong negative inter-metabolite correlation between citrate and creatinine and positive between 2-oxoglutarate and creatinine (ii),(iii) and strong positive intra-metabolite correlations for creatinine (iv) and citrate (v). Colour scale indicates the level of Pearson correlation.

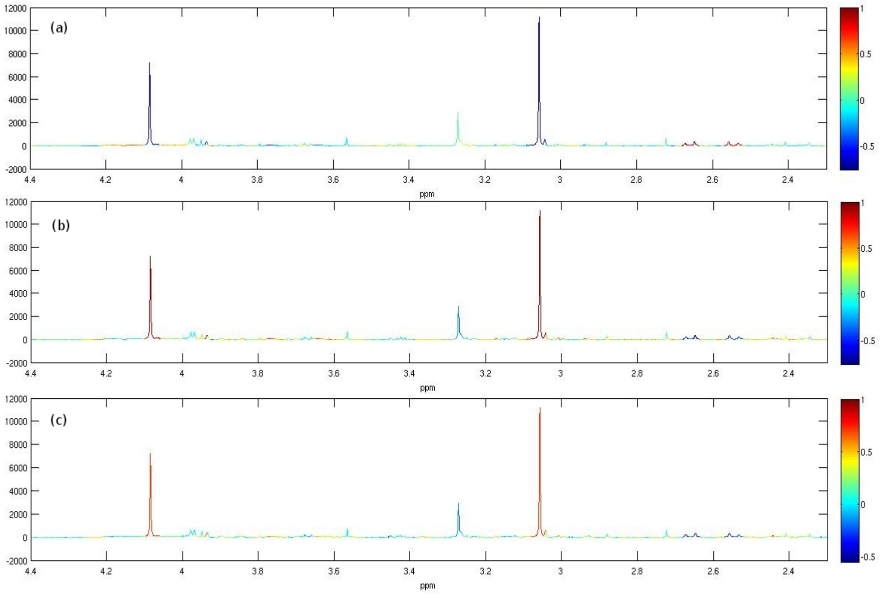


Figure 2.8: Pairwise correlation coefficients mapped as a colour code onto the mean spectrum. Correlations to (a) citrate 2.65ppm, (b) creatinine 4.08ppm, (c) 2-oxoglutarate 2.44ppm.

Chapter 3

Modelling pH-induced changes in ^{1}H-NMR chemical shifts

3.1 Background

Uncontrolled variation in chemical shift position of NMR resonances is a major challenge in statistical analysis of complex mixtures and greatly complicates signal deconvolution and metabolite identification. Characterising this variation is an ongoing problem within metabolomics, and there are several strategies available such as potentiometric titrimetry, UV spectrophotrometry or capillary zone electrophoresis can be used to experimentally measure pK$_a$ values, but NMR is by far the most common within metabolomics [80, 81, 82, 83]. NMR pH titration involves measuring the chemical shift of an analyte at varying pH values, resulting in titration curves mapping the relationship between chemical shift and pH.

There exists much literature detailing the derivation of ^{1}H-NMR peak shift parameters of particular metabolites [84, 85, 86, 87]. Knowing these parameters is very useful in aiding peak deconvolution methods such as BATMAN [60] and metabolite identification in addition to furthering fundamental chemical knowledge. Also, by enabling improved deconvolution more reliable chemical shift data will become available allowing us to model the parameters of these complex peak patterns, continually improving our characterisation of the data. Improved estimates are also highly beneficial in producing realistic simulation techniques. For example, MetAssimulo (described in Chapter 2) uses this information to model realistic peak shifts depending on the desired mixture pH [65]. With this motivation in mind, the positional variation due to pH changes of several peaks is modelled with the aim of providing realistic estimates of the parameters involved.

3.1.1 Modelling Titration Curves

Protonation is the fundamental reaction of adding a proton (H$^+$) to a molecule. The protonated and deprotonated species of a particular molecule will have different chemical shifts and a molecule may have multiple sites where it is possible to bind this extra proton. The simplest case is when a molecule A has a single protonation site - the monoprotic case. In the fast-exchange regime the rate constant (speed of reaction) of the exchange process between the different states or species is much larger than the difference in chemical shift between the states. In this setting, a common resonance, δ^{obsd}, of two species (the protonated and unprotonated versions of the molecule: HA^+ and A) can be represented by the weighted average of δ_A and δ_{HA}, the limiting chemical shifts [84]:

$$\delta^{obsd} = \delta_A \frac{[A]}{[A] + [HA^+]} + \delta_{HA} \frac{[HA^+]}{[A] + [HA^+]} \tag{3.1}$$

where $[A]$ denotes the concentration of molecule A. Using the definitions of the acid dissociation constant, pK$_a$, and pH given by Equations 3.2-3.4:

$$pH = -log_{10}[H^+] \tag{3.2}$$

$$pKa = -log_{10}K_a \tag{3.3}$$

$$K_a = \frac{[H^+][A]}{[HA^+]} \tag{3.4}$$

and substituting into Equation 3.1 the Henderson Hasselbach Equation is obtained:

$$\delta^{obsd} = \frac{10^{pH-pKa}\delta_A + \delta_{HA}}{1 + 10^{pH-pKa}} \tag{3.5}$$

The pK$_a$ of an acid explains how acidic a particular hydrogen atom in a molecule is, whereas pH is a measure of how acidic a solution is. pK$_a$ is the pH at which it is exactly half dissociated [88]: the inflection point of the titration curve (see Figure 3.1 for a monoprotic example). It is at this point that the concentrations of the acid and base conjugates are in equilibrium. The acid/base chemical shift limits can be seen as the asymptotes of the curve at the two extreme pH values.

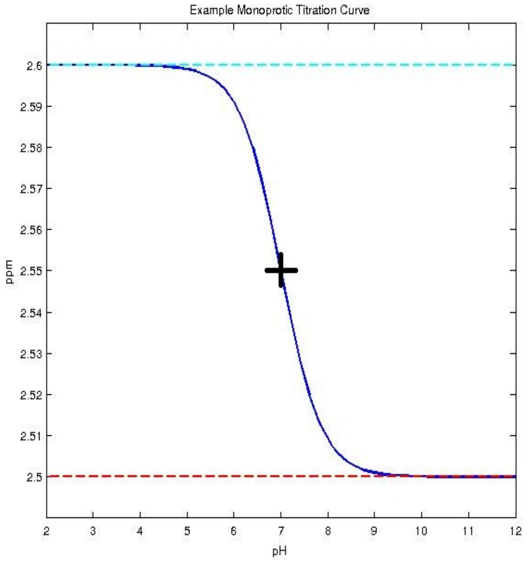


Figure 3.1: An example monoprotic titration curve modeled by the Henderson Hasselbach Equation is shown in blue. The pK_a value is 7 and the inflection point is indicated with a black cross. The acid and base asymptotic limits are given by cyan and red dashed lines respectively.

There are several approaches to modelling ^{1}H-NMR titration curves. For example, the modified Hill Equation:

$$\delta^{obsd} = \delta_A[1 + 10^{n(pKa-pH)}]^{-1} + \delta_{HA}\{1 - [1 + 10^{n(pKa-pH)}]^{-1}\} \tag{3.6}$$

first presented by Markley [89], is often used as an alternative to the Henderson Hasselbach when the titration curve appears irregular or deviates from a single ionization equilibrium [90, 87]. However, n, the Hill coefficient, can be difficult to interpret. Equally, one can use the system of dissociation equilibrium equations exploiting the stoichiometry of the protonation process [91, 92, 93].

When dealing with polyprotic cases, i.e. single titration curves with multiple equilibria, the situation becomes more complex. Each equilibrium has its own pK$_a$ value and acid/base limits shared with the adjacent equilibria, increasing the number of unknown variables. It can be useful to investigate the equilibria at the microscopic level by examining the relations between microconstants and macroconstants (pK$_a$ values) [94, 95, 96, 84]. However, since the number of unknowns in the system increases for higher numbers of sites, the titration curves are split into separate curves each containing a single equilibrium if possible. If adjacent equilibria are a significant distance apart, i.e. at least four covalent bonds which corresponds to a difference in pK$_a$ value of about 3 or more, the equilibria can be separated and modelled as individual Henderson Hasselbachs [84]. Indeed, a common technique is to model polyprotic systems using a mixture of Henderson Hasselbach Equations [94, 97].

An alternative to a Henderson Hasselbach mixture is an equation described by Szakacs (Equation 3.7) [98]. This approach generalises Equation 3.1 for an arbitrary number of protonation sites (q) and reduces to the Henderson Hasselbach Equation for $q = 1$:

$$\delta^{obsd} = \frac{\delta_A + \Sigma_{i=1}^{q} \delta_{H_iA} 10^{(\Sigma_{j=q-i+1}^{q} pKa_j) - ipH}}{1 + \Sigma_{k=1}^{q} 10^{(\Sigma_{l=q-k+1}^{q} pKa_l) - kpH}} \tag{3.7}$$

The values $(pKa_1, ..., pKa_q)$ denote the pK$_a$ values associated with each of the q equilibrium points and $(\delta_A, \delta_{H_1A}, ..., \delta_{H_qA})$ are the relevant chemical shift limits. An example titration curve for the polyprotic model with $q = 2$ is shown in Figure 3.2.

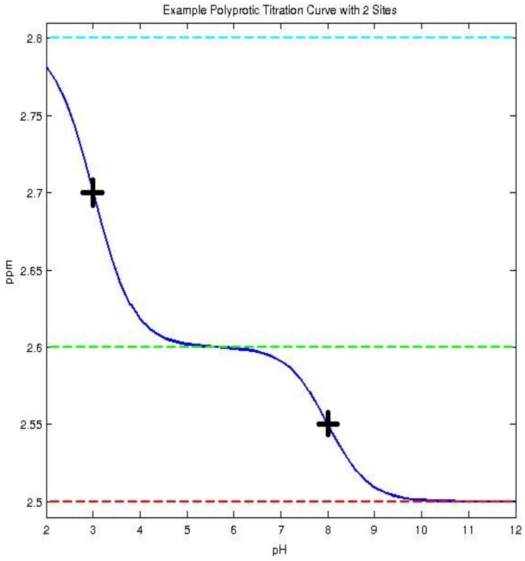


Figure 3.2: An example polyprotic titration curve model for a two site molecule is shown in blue. The pKa values are 3 and 8 with the inflection points indicated by black crosses. The acid/base asymptotic limits are given by cyan, green and red dashed lines.

3.1.2 Model Selection Criteria

It would be useful to be able to predict the number of sites the metabolite has using the titration curves. To this end, a reliable method of selecting the (polyprotic) model with the correct number of sites is required. There are several strategies available for model selection and here some of them are examined.

Akaike Information Criterion

The Akaike Information Criterion (AIC) was developed by Akaike in 1974 as a measure of the goodness of fit of models [99]. It is calculated using the equation:

$$\mathrm{AIC} = -2 \cdot \ln L + 2 \cdot k \tag{3.8}$$

where L is the likelihood evaluated at the maximum and k is the number of free parameters within the model. The AIC aims to strike a balance between the accuracy and complexity of the model by penalising the likelihood by the number of parameters of the model and as such the smallest value indicates the best model. AIC may be preferred if the main application of the model is prediction, whereas in cases where the primary goal is description (i.e. building a model incorporating the most influential variables) one may opt for the Bayesian Information Criterion.

Bayesian Information Criterion

Similarly constructed to the AIC, the Bayesian Information Criterion (BIC) developed by Schwarz in 1978 [100] is:

$$-2 \cdot \ln p(\mathbf{y}|k) \approx \mathrm{BIC} = -2 \cdot \ln L + k \cdot \ln(n) \tag{3.9}$$

where $\mathbf{y} = (y_1, ..., y_n)$ is the observed data and n is the sample size. Again, the smallest value indicates the best model as one attempts to select the model corresponding to the highest posterior probability. BIC uses the same goodness-of-fit term as AIC, but it has been argued that the BIC penalises over-fitting more effectively than the AIC measure [101]. Under certain settings the BIC is also roughly equivalent to selection using Bayes Factors, so has appeal in problems where priors are hard to define accurately.

Deviance Information Criterion

The Deviance Information Criterion (DIC) is described as a hierarchical modelling generalisation of the AIC and BIC and is defined as:

$$\mathrm{DIC} = p_D + \bar{D} \tag{3.10}$$

$$p_D = \bar{D} - D(\bar{\theta}) \tag{3.11}$$

$$\bar{D} = \mathbf{E}_\theta[D(\theta)] \tag{3.12}$$

$$D(\theta) = -2\log(p(y|\theta)) + C \tag{3.13}$$

where C is a constant that cancels in all calculations comparing models. $D(\theta)$ is defined as the deviance and the expectation of this quantity, $\bar{D}$, gives a measure of goodness of fit. p_D is the number of effective parameters of the model, whilst $\bar{\theta}$ is the posterior expectation of θ [55, 102]. Here, again the smallest value is preferred for the best model. One major advantage of DIC is that it is easily computed from MCMC output. However, the observed data is used in constructing the posterior distribution as well as evaluating the candidate models so DIC may have a tendency to promote over-fitting.

Log Pseudo Marginal Likelihood

The Log Pseudo Marginal Likelihood (LPML) is a summary statistic calculated from the Conditional Posterior Ordinate (CPO) values using the harmonic mean [103, 104]:

$$\text{LPML} = \Sigma_{i=1}^{n} \log(\text{CPO}_i) \tag{3.14}$$

where

$$\text{CPO}_i = p(y_i|\mathbf{y_{[i]}}) = \int p(y_i|\theta)p(\theta|\mathbf{y_{[i]}})d\theta \tag{3.15}$$

and $\mathbf{y_{[i]}}$ denotes $\mathbf{y}$ with the i-th value eliminated. The CPO can also be described as the 'leave-one-out cross-validation predictive density' and has been proposed as an alternative to model selection techniques that employ a 'double-use' of the data. The CPO$_i$ values can be estimated from MCMC output using the following formula [105]:

$$\hat{\text{CPO}}_i = \frac{1}{T^{-1}\Sigma_{t=1}^{T}p(y_i|\theta^{(t)}) - 1} \tag{3.16}$$

where $\theta^{(t)}$ denotes MCMC iterations from $t = 1, .., T$. CPO can be considered as the posterior probability of observing y_i when the model is fitted to all the data with y_i omitted. Small values of CPO$_i$ are often used to detect outliers, and larger values of the LPML indicate a better model fit. A ratio of pseudo marginal likelihoods is often considered as a surrogate for the Bayes Factor, which can often be difficult to calculate exactly for complex models, and may be described as the ‘pseudo Bayes factor’.

Bayes Factors

Bayes Factors are a standard method for Bayesian model comparison defined as the ratio of the marginal likelihood under one model, say M_1, to the marginal likelihood under the second model M_2 [55] :

$$\frac{p(\mathbf{y}|M_2)}{p(\mathbf{y}|M_1)} \tag{3.17}$$

Values greater than 10 are considered substantial evidence supporting M_1 and by similarly values less than 1/10 support M_2. Bayes Factors are considered to penalise model complexity more effectively than the other methods outlined here and thus should have less tendency to choose over-fitted models.

$p(\mathbf{y}|M_i)$ can be estimated from MCMC output in a variety of ways [101], but here a method involving importance sampling is employed. The importance sampling estimate centres around the identity:

$$1 = \int \frac{p(\mathbf{y})p(\theta|\mathbf{y})}{p(\mathbf{y}|\theta)p(\theta)} g(\theta) d\theta \tag{3.18}$$

which gives the approximation:

$$\hat{p}(\mathbf{y})^{-1} = T^{-1}\Sigma_{t=1}^{T} \frac{g^{(t)}}{p(\mathbf{y}|\theta^{(t)})p(\theta^{(t)})} \tag{3.19}$$

where $\theta^{(t)}$ denotes MCMC iterations from $t = 1, .., T$, and $g^{(t)}$ denotes T samples from the importance sampling function g. An importance sampling function taken from Newton & Raftery based on combining samples from both the prior and posterior [106] is utilised:

$$g(\theta) = \delta p(\theta) + (1 - \delta)p(\theta|\mathbf{y}) \tag{3.20}$$

with $0 < \delta < 1$. As recommended by the literature $\delta = 0.05$. From the same source, a synthetic estimator is used that avoids sampling from the prior density, with the algorithmic form detailed in the Appendix.

Likelihood Ratio Test

The Likelihood Ratio Test is a common frequentist method for model comparison, but mentioned here for contrast. It is assumed that the ratio of the likelihoods of two models (M_1 =null, M_2 =alternative) has a χ^2 distribution with degrees of freedom $df_2 - df_1$ where df_i is the number of parameters of model M_i and θ^* is the value at which the likelihood attains its maximum:

$$L = \frac{p(\mathbf{y}|\theta^*, M_1)}{p(\mathbf{y}|\theta^*, M_2)} \tag{3.21}$$

$$-2\ln(L) \sim \chi^2_{df_2 - df1} \tag{3.22}$$

There is no consideration to Bayesian prior probabilities here, but it could be used to compare the frequentist version of a model. In contrast to the Bayesian perspective where we might ask to what extent does the data support one model over the other, here we ask whether the maximum likelihood of the alternative hypothesis is large enough to reject the null hypothesis. Since Bayes Factors consider the marginal likelihoods (likelihood averaged over all possible parameter values) they, unlike the frequentist likelihood comparison, do not rely on a single set of parameter values. It could be said that the strength of evidence required to advocate a more complex model is greater in Bayesian analysis than frequentist since the analytical paradigm incorporates parameter uncertainty in addition to the estimation uncertainty present in frequentist analysis[107].

3.2 Data

A large urine sample from five different individuals was collected and pooled to obtain an average representative human urine sample. In order to evaluate if altering the pH in the urine samples resulted in a change in ionic concentration, the main metal ion concentrations in urine (Ca2+, Mg2+, Na+ and K+) were measured with ion selective electrodes following the adjustment of human urine with acid (HCl) and base (NaOH). Some changes in the metal ion concentrations occurred at the extremes of pH and this effect was more pronounced for the divalent ions Ca2+ and Mg2+, so it was felt important to remove these ions by treating the urine with chelex resin.

Treating the urine with chelex resin did not significantly alter the metabolite composition of human urine as seen by ^{1}H NMR spectra, but both Ca2+ and Mg2+ ion concentrations were reduced. It must be noted that neither Na+ nor K+ ions were removed by the chelex resin, and in fact Na+ ions were slightly increased as they were displaced from the chelex resin by the divalent ions. However, it was noted from the literature that the main ionic contributors to NMR peak shifts were the divalent ions Ca2+ and Mg2+.

Spectra were acquired on a Bruker Avance DRX600 NMR spectrometer (Bruker BioSpin, Rheinstetten, Germany), with ^{1}H frequency of 600 MHz. Samples were introduced with an automatic sampler. A one-dimensional NOESY sequence was used for water suppression; data were acquired into 64K data points over a spectral width of 12 KHz, with 8 dummy scans and 64 scans per sample. Spectra were processed in iNMR 3.4 (Nucleomatica, Molfetta, Italy). Fourier transform of the free-induction decay was applied with a line broadening of 0.5Hz. Spectra were manually phased and automated first order baseline correction was applied. Metabolites were assigned using the Chenomx NMR Suite 5.1 (Chenomx, Inc., Edmonton, Alberta, Canada) relative to 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS).

Metabolite peak positions from the different samples were obtained using in-house Matlab scripts, and appropriate chemical shifts were calculated for multiplets relative to DSS for the metabolites shown in Figure 3.3 with the number of sites given in Table 3.1. All samples were prepared and processed by Gregory Tredwell (Imperial College London, UK).

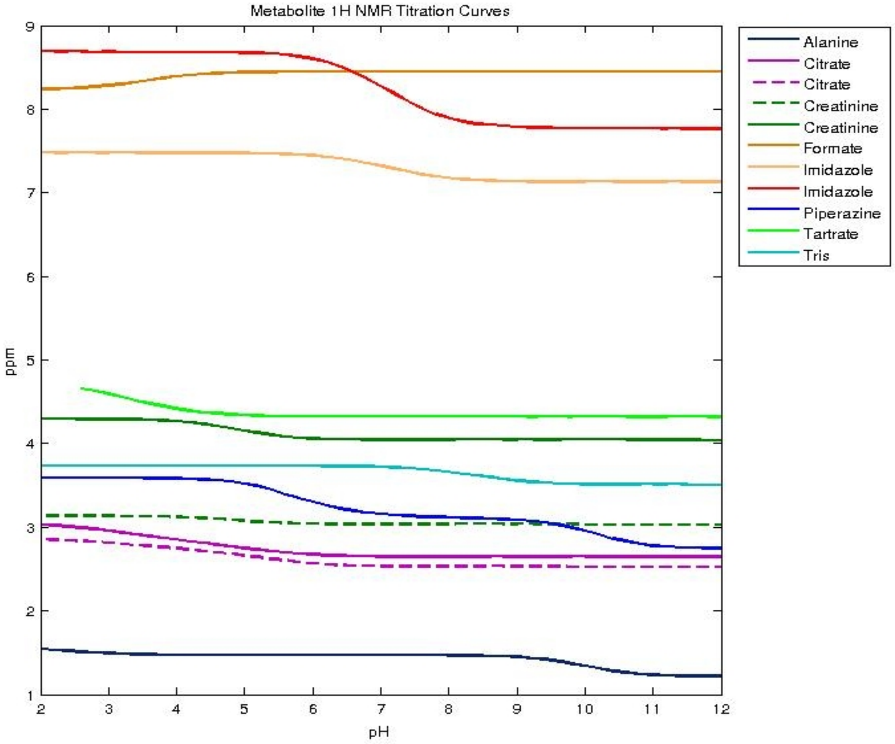


Figure 3.3: Metabolite ^{1}H NMR titration curves showing varying effects of pH on several different labelled metabolite resonances.

Metabolite	No. of Sites
Formate	1
Tris	1
Imidazole 1	1
Imidazole 2	1
Alanine	2
Creatinine 1	2
Creatinine 2	2
Piperazine	2
Tartrate	2
Citrate 1	3
Citrate 2	3

Table 3.1: Number of protonation sites for each metabolite studied. Where multiple resonances are used for a single metabolite these are labeled numerically.

3.3 Model Construction

For a molecule with q protonation sites a non-linear regression model is constructed using Equation 3.7. Treating the observed chemical shift for n different pH levels as our response variable $\mathbf{y} = (y_1, ..., y_n)$, gives:

$$\begin{aligned} \mathbf{y}|\mu, \sigma^2 &\sim N(\mu(\mathbf{pH}, \mathbf{pKa}, \delta_{\mathbf{A}}, q), \sigma^2) \\ \mu_i = \frac{\delta_A + \Sigma_{m=1}^{q} \delta_{H_m A} 10^{(\Sigma_{j=q-m+1}^{q} pKa_j) - mpH_i}}{1 + \Sigma_{k=1}^{q} 10^{(\Sigma_{l=q-k+1}^{q} pKa_l) - kpH_i}} & \\ \sigma^2 &\sim IG(\alpha/2, \beta/2) \\ pKa_t &\sim U(0, 14) \text{ for } t = 1, ..., q \\ \delta_A &\sim U(0, 10) \\ \delta_{H_t A} &\sim U(0, 10) \text{ for } t = 1, ..., q \end{aligned} \tag{3.23}$$

where the covariates consist of known $\mathbf{pH} = (pH_1, ..., pH_n)$ and fixed q whilst the q pK$_a$ values, $\mathbf{pKa} = (pKa_1, ..., pKa_q)$, and $q+1$ acid/base limits, $\delta_{\mathbf{A}} = (\delta_A, \delta_{H_1 A}, ..., \delta_{H_q A})$, for each equilibria are assumed unknown. Uniform priors are assigned to each pKa_i and $\delta_A, \delta_{H_i A}$ as physiologically speaking there should be no prior preference for particular values, and an Inverse-Gamma to σ^2 with parameters $(\alpha/2, \beta/2)$ for the convenience of conjugacy. The **pKa** values are restricted to ascending order within the interval [0,14] given that is the range of the pH scale, whilst the $\delta_{\mathbf{A}}$ values are allowed to vary freely over the NMR ppm scale [0,10]. Is it assumed that there are no additional spectral effects.

3.3.1 Model Tuning and Convergence

Let us investigate model convergence for Imidazole 1, one of the chemical shifts of a one site molecule using the one site model.

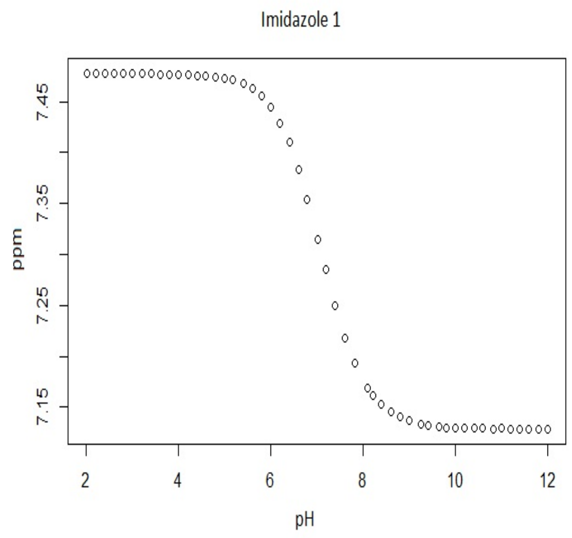


Figure 3.4: Metabolite ^{1}H NMR titration curve for Imidazole 1 (a one site molecule with pKa = 6.95)

The Imidazole curve has a single inflexion point at $pKa = 6.95$ and the chemical shift varies between 7.13 and 7.48 for this titration experiment. Figure 3.5 shows a trace plot and density of the pKa sampling for a run of 100,000 iterations of four chains with a burn-in of 100,000 and thinning parameter of 10 in *rjags*.

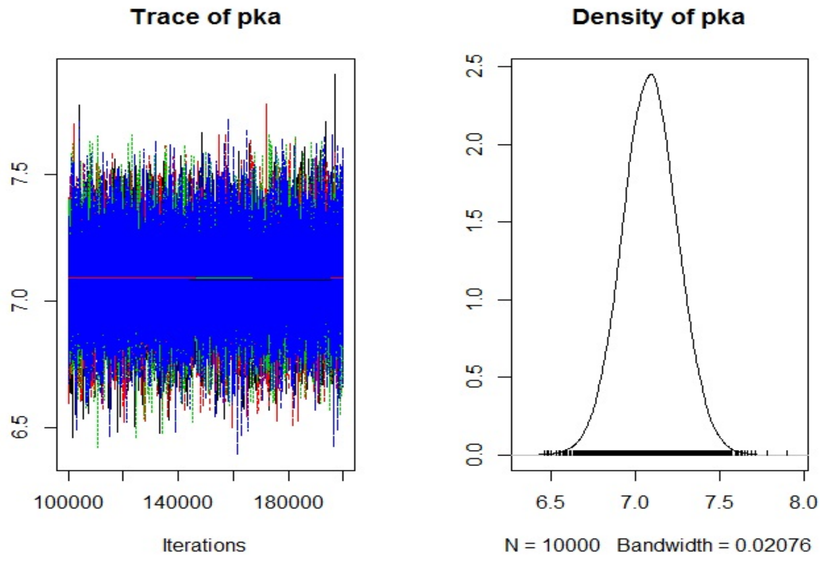


Figure 3.5: MCMC trace and density plot for pKa of Imidazole using 1 site model. 4 chains run with 100,000 iterations after a burn-in of 100,000 and thinning parameter of 10.

The trace shows that the sampler quickly converges towards the stationary distribution and the density looks nicely normal with a mean of 7.09. Figure 3.6 shows the trace and density for the acid and base limits. There does not seem to be strong convergence in the trace plots, indeed the separate chains are not particularly correlated and the densities are multimodal. The acid and base limit sampling is highly correlated and the model has difficulty with estimating these variables.

Re-running with many more iterations, 500,000 iterations after a burn-in period of 500,000 iterations and increasing the thinning parameter to 100 does not improve the situation as shown in Figure 3.7.

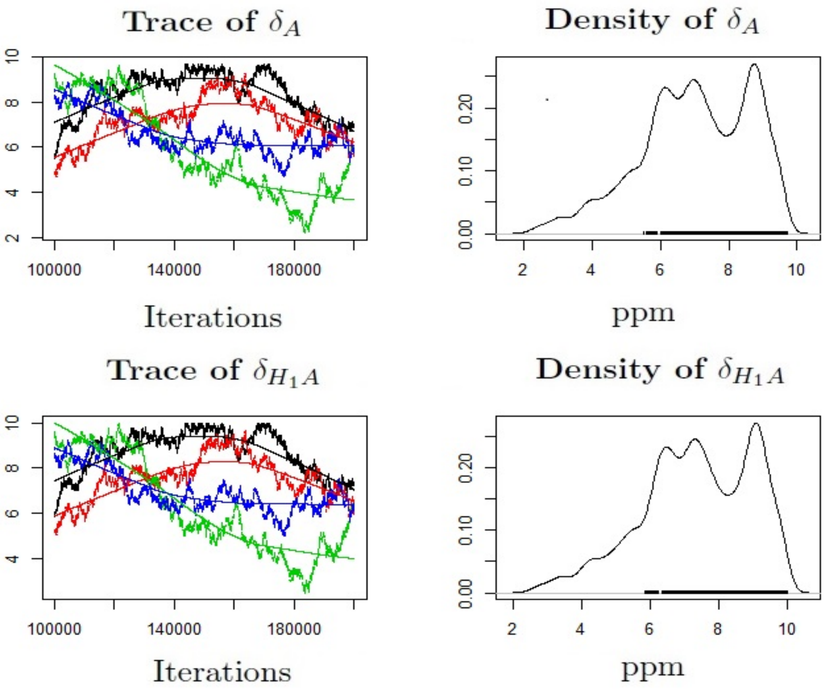


Figure 3.6: MCMC trace and density plot for acid and base limits of Imidazole using 1 site model. 4 chains run with 100,000 iterations after a burn-in of 100,000 and thinning parameter of 10.

A stronger prior on both acid and base limits was investigated to help improve convergence: a truncated Normal distribution centred at the midpoint between 0 and 10 with variance of 1:

$$\delta_A \sim N(5,1)I(0,10)$$

$$\delta_{HA} \sim N(5,1)I(0,10) \tag{3.24}$$

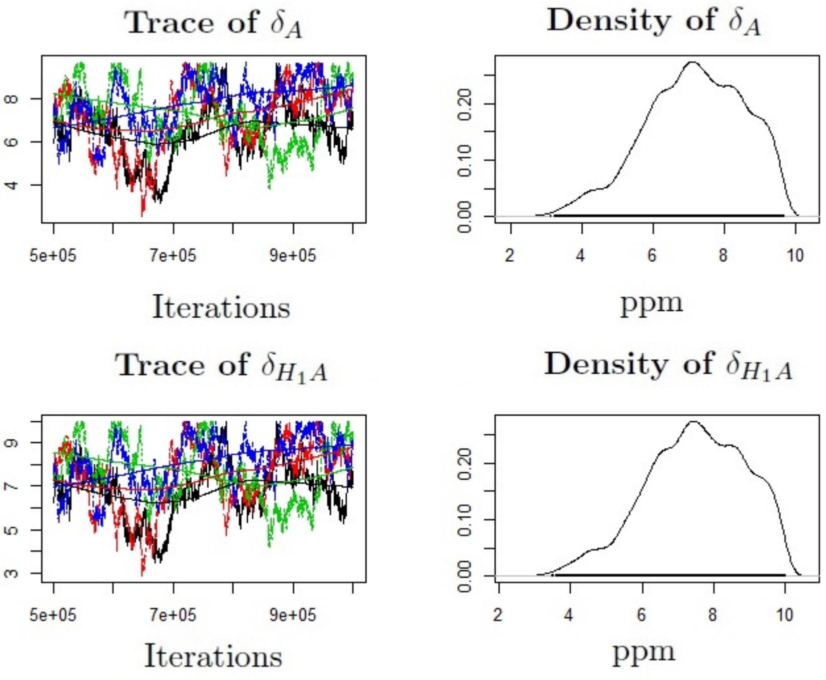


Figure 3.7: MCMC trace and density plot for acid and base limits of Imidazole using 1 site model. 4 chains run with 500,000 iterations after a burn-in of 500,000 and thinning parameter of 100.

although physiologically speaking we should not prefer any particular ppm region to another.

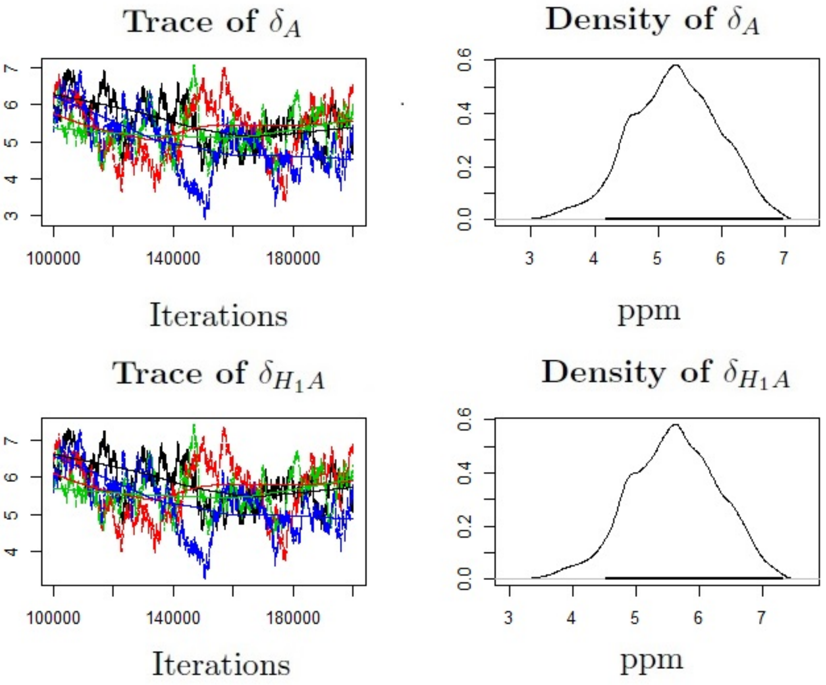


Figure 3.8: MCMC trace and density plot for acid and base limits of Imidazole using 1 site model with prior N(5,1)I(0,10). Multiple chains run for 100,000 iterations after a burn-in of 100,000 and thinning parameter of 10.

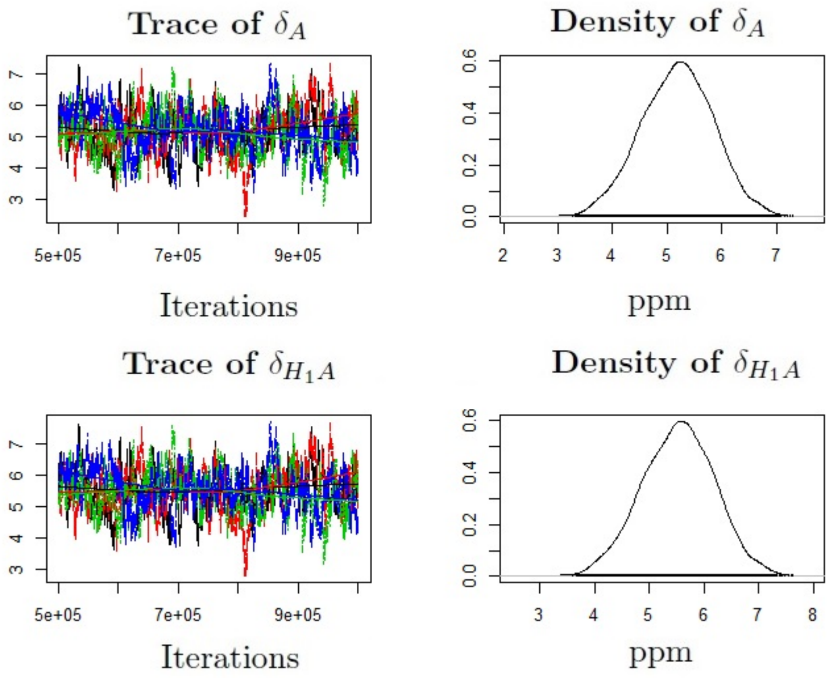


Figure 3.9: MCMC trace and density plot for acid and base limits of Imidazole using 1 site model with prior N(5,1)I(0,10). Multiple chains run for 500,000 iterations after a burn-in of 500,000 and thinning parameter of 50.

Figure 3.8 shows traces and density for multiple chains run with the stronger prior. Although the density is now uni-modal, the distribution is quite strongly influenced by the mean of the prior distribution. Increasing the thinning parameter to 50 and running longer burn-in (500,000 iterations) results in a better trace plot although the posterior distribution is skewed toward the prior mean shown in Figure 3.9.

Since the prior distribution now seems too strong given the skew towards 5, we increase the variance of the prior:

$$\delta_A \sim N(5,2)I(0,10)$$
$$\delta_{HA} \sim N(5,2)I(0,10) \tag{3.25}$$

The resulting sample density is still too strongly influenced by the prior mean seen in Figure 3.10.

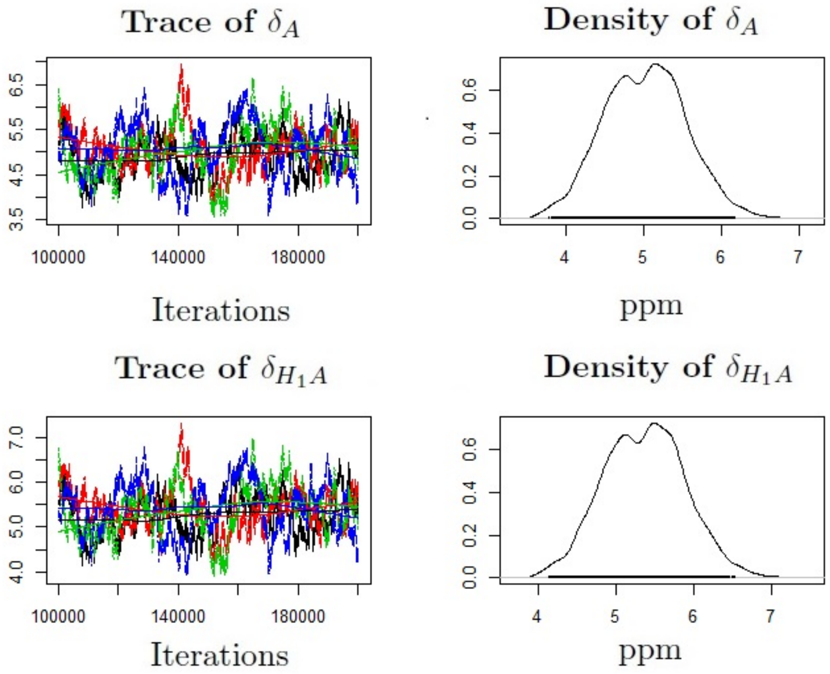


Figure 3.10: MCMC trace and density plot for acid and base limits of Imidazole using 1 site model with prior N(5,2)I(0,10). 4 chains run for 100,000 iterations after a burn-in of 100,000 and thinning parameter of 10.

The analysis is very sensitive to prior specification as it does not appear that there is enough information in the data to estimate all the parameters. Analysis will be continued using the original uniform prior distribution on the acid/base limits.

Let us examine the convergence of the two site model for a two site molecule chemical shift: Piperazine, where the true pKa values are 5.56 and 9.83.

Figure 3.11 shows nice convergence in the trace plots with unimodal posterior distributions centred close to the true values. Looking at Figure 3.12 we have a similar indentifiability issue as for the one site model.

So again the stronger prior, N(5,1)I(0,10), is used with results shown in Figure 3.13. Similarly for the one site model, this does slightly improve the performance of the sampling although mixing does not seem optimal. Despite the issue of acid and base limit parameter identifiability I explore the model further to investigate whether or not we can still garner useful information about the molecular structure of the metabolite from the titration curves.

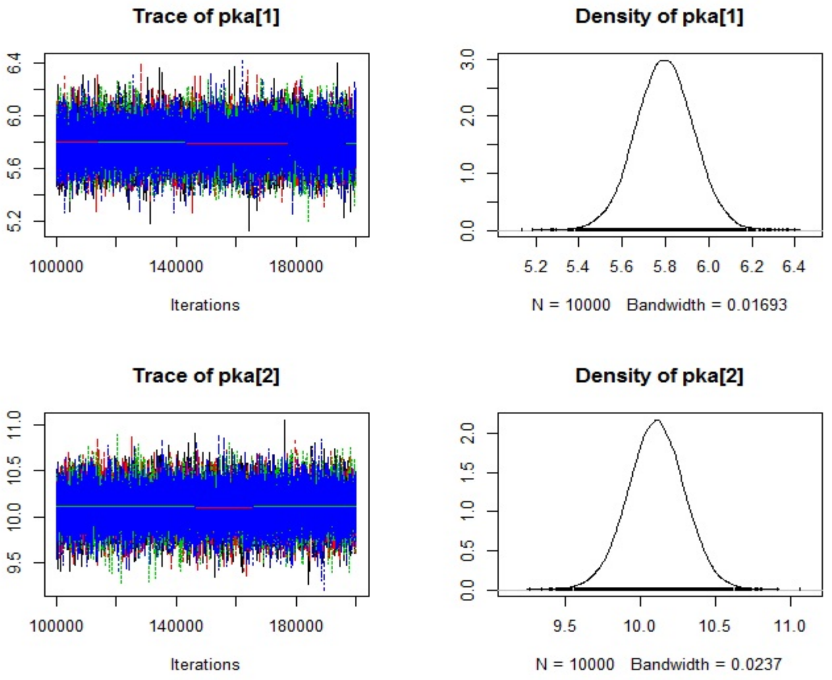
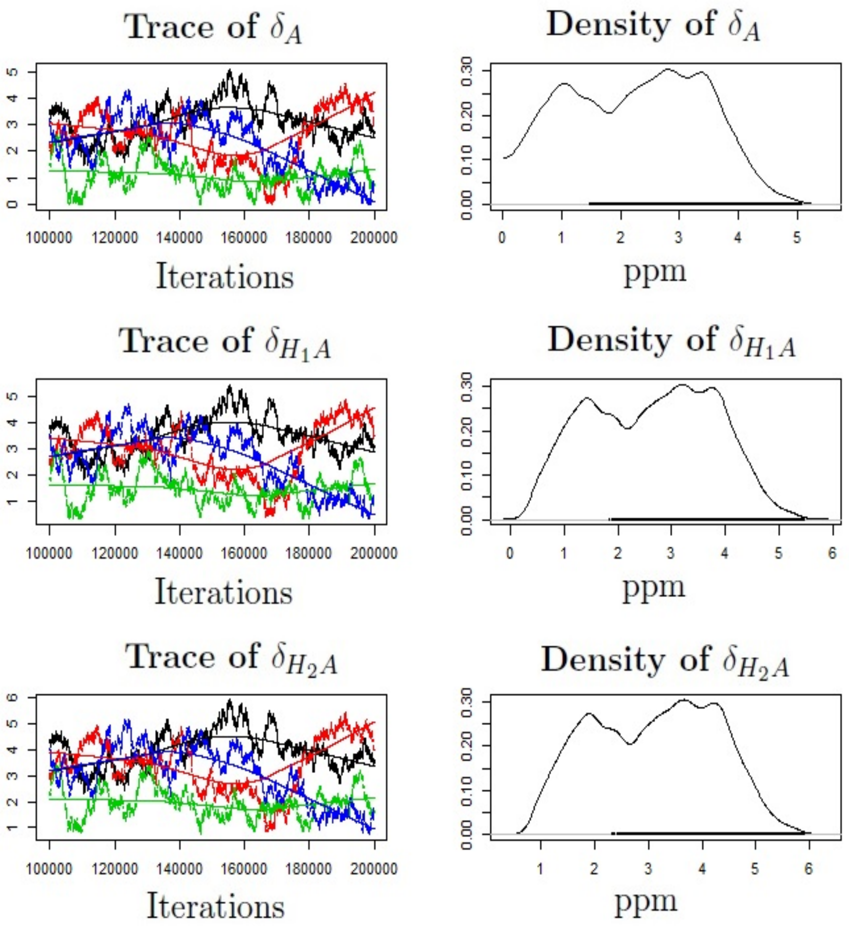


Figure 3.11: MCMC trace and density plot for pka of Piperazine using 2 site model. 4 chains run for 100,000 iterations after a burn-in of 100,000 and thinning parameter of 10.

3.3.2 Model Selection

Figure 3.14 shows our fit for three different metabolites: Formate, Creatinine and Citrate which have 1, 2 and 3 sites respectively.

The plots show generally good fits for the data, which is contained within the 95% credible intervals. The fits are also compared with a non-linear least-squares fitting algorithm in MATLAB called *nlinfit* in Figure 3.15. Posterior estimates and variances of pK$_a$ values were calculated using JAGS (Just Another Gibbs Sampler) via the R package *rjags* [59] and are shown in Table 3.2 alongside *nlinfit* with mean squared errors. It should be noted that frequentist and Bayesian measures of uncertainty (i.e. mean square error vs posterior parameter variance) are conceptually different and therefore not directly comparable but I have included the *nlinfit* MSE here for illustration to show that the titration curves *can* be very closely fit by the chosen function. A number of different model selection criteria were investigated, focusing on the 1, 2 and 3 site models.


Trace of δ_A

Density of δ_A

Trace of δ_{H_1A}

Density of δ_{H_1A}

Trace of δ_{H_2A}

Density of δ_{H_2A}

Figure 3.12: MCMC trace and density plot for acid and base limits of Piperazine using 2 site model. 4 chains run for 100,000 iterations after a burn-in of 100,000 and thinning parameter of 10.

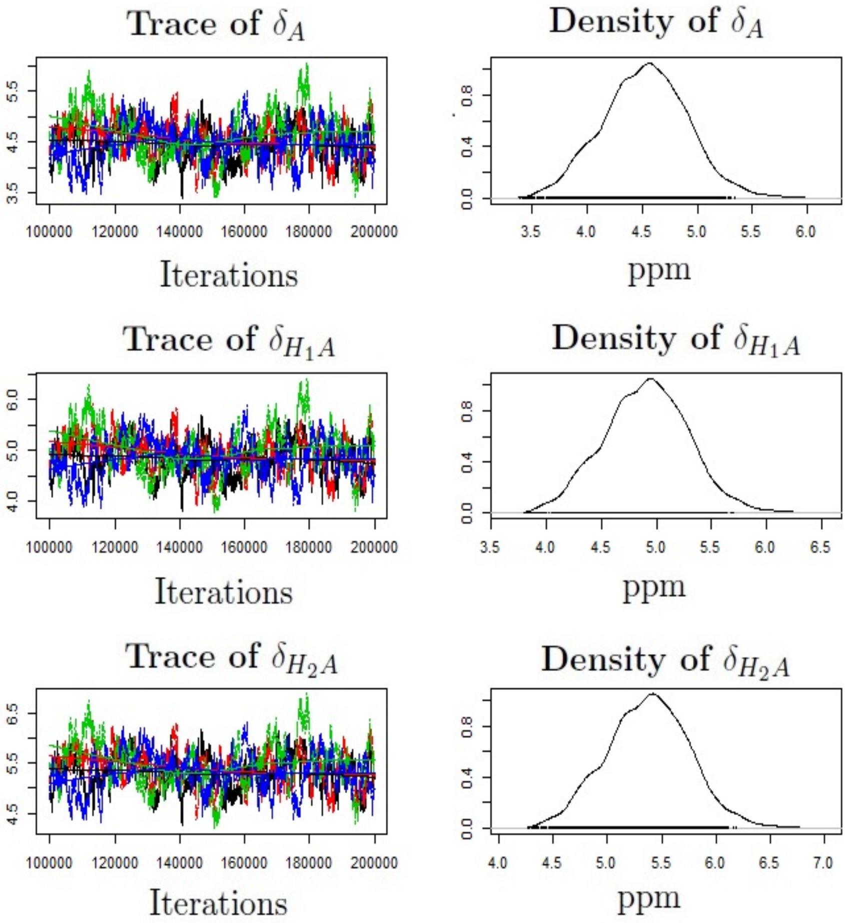


Figure 3.13: MCMC trace and density plot for acid and base limits of Piperazine using 2 site model with stronger prior N(5,2)I(0,10). 4 chains run for 100,000 iterations after a burn-in of 100,000 and thinning parameter of 10.

Metabolite	Actual pKa	Posterior pKa Mean	Posterior pKa Variance	*nlinfit* pKa Estimate	*nlinfit* MSE
Formate	3.77	3.51	0.020	3.56	2.4E-6
Tris	8.30	8.36	0.016	8.36	1.1E-4
Imidazole 1	6.95	7.09	0.016	7.09	8.0E-5
Imidazole 2	6.95	7.10	0.012	7.10	6.5E-5
Alanine	2.35	2.87	6.36	2.39	8.7E-4
	9.69	10.68	1.41	9.98	2.0E-5
Creatinine 1	4.84	2.75	2.34	4.89	8.4E-6
	9.2	8.76	4.30	11.30	0.02
Creatinine 2	4.84	3.06	2.14	4.89	5.0E-6
	9.2	9.17	4.21	12.10	0.24
Piperazine	5.56	5.80	0.13	5.79	9.2E-5
	9.83	10.11	0.19	10.10	1.8E-4
Tartrate	2.98	1.56	1.16	2.90	1.3E-4
	4.34	5.35	3.93	3.97	7.3E-5
Citrate 1	3.09	1.06	1.25	2.94	2.0E-5
	4.75	3.69	3.00	4.29	7.6E-5
	5.41	8.3	4.32	5.48	2.7E-5
Citrate 2	3.09	3.53	5.14	2.94	0.001
	4.75	7.08	36.25	6.02	0.003
	5.41	13.47	0.73	5.51	2.5E-4

Table 3.2: Actual pKa values vs Posterior Mean pKa Estimates and Variance from MCMC simulation using *rjags* package and MATLAB function *nlinfit* pKa Estimates with mean squared error (MSE) values.

Metabolite	Actual Sites	1 Site Model	2 Site Model	3 Site Model
Formate	1	-671.8	-679.3	**-724.9**
Tris	1	-449.6	**-610.0**	-609.8
Imidazole 1	1	-422.7	-563.6	**-618.6**
Imidazole 2	1	-334.7	-488.0	**-517.4**
Alanine	2	-275.3	**-532.1**	-528.2
Creatinine 1	2	-566.9	-566.4	**-664.5**
Creatinine 2	2	-561.3	-557.9	**-603.0**
Piperazine	2	-78.4	-390.8	**-482.1**
Tartrate	2	-382.3	-654.7	**-662.1**
Citrate 1	3	-247.3	-444.1	**-738.0**
Citrate 2	3	-269.5	-468.6	**-590.1**

Table 3.3: Akaike Information Criterion values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Smallest values are shown in bold, indicating chosen model.

Metabolite	Actual Sites	1 Site Model	2 Site Model	3 Site Model
Formate	1	-664	-668	**-709**
Tris	1	-442	**-598**	-594
Imidazole 1	1	-415	-552	**-603**
Imidazole 2	1	-327	-476	**-502**
Alanine	2	-268	**-521**	-513
Creatinine 1	2	-559	-555	**-649**
Creatinine 2	2	-554	-546	**-588**
Piperazine	2	-71	-379	**-467**
Tartrate	2	-375	-644	**-647**
Citrate 1	3	-240	-432	**-723**
Citrate 2	3	-262	-457	**-575**

Table 3.4: Bayesian Information Criterion values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Smallest values are shown in bold, indicating chosen model.

Metabolite	Actual Sites	1 Site Model	2 Site Model	3 Site Model
Formate	1	-170.4	**-177.3**	-175.9
Tris	1	-170.6	**-177.2**	-175.6
Imidazole 1	1	-170.6	**-177.0**	-175.7
Imidazole 2	1	-169.4	**-175.8**	-174.2
Alanine	2	-166.1	**-174.5**	**-174.5**
Creatinine 1	2	-166.7	**-175.7**	-173.9
Creatinine 2	2	-170.8	**-177.0**	-176.0
Piperazine	2	-74.6	**-173.0**	-170.7
Tartrate	2	-155.7	**-162.6**	-162.3
Citrate 1	3	-162.4	**-171.1**	170.3
Citrate 2	3	-165.2	**-173.1**	-172.2

Table 3.5: Deviance Information Criterion values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Smallest values are shown in bold, indicating chosen model.

Metabolite	Actual No. of Params.	1 Site Model p_D	2 Site Model p_D	3 Site Model p_D
Formate	4	4.2	4.8	5.3
Tris	4	4.2	5.0	5.5
Imidazole 1	4	4.1	4.9	5.4
Imidazole 2	4	4.128	5.0	5.5
Alanine	6	4.2	5.4	5.7
Creatinine 1	6	5.3	5.3	5.9
Creatinine 2	6	4.1	4.9	5.4
Piperazine	6	5.1	6.3	7.0
Tartrate	6	4.6	5.1	5.2
Citrate 1	8	4.7	5.9	6.3
Citrate 2	8	4.6	5.6	6.0

Table 3.6: Actual number of free parameters for appropriate model and p_D values (estimated number of effective parameters) for each model.

Metabolite	Actual Sites	1 Site Model	2 Site Model	3 Site Model
Formate	1	86.97	**90.50**	89.92
Tris	1	87.07	**90.57**	90.17
Imidazole 1	1	86.95	**90.50**	90.11
Imidazole 2	1	86.50	**89.83**	89.43
Alanine	2	84.65	**89.01**	88.97
Creatinine 1	2	85.21	89.16	**89.85**
Creatinine 2	2	87.08	**90.76**	90.18
Piperazine	2	38.50	**89.03**	88.20
Tartrate	2	79.50	**83.21**	82.77
Citrate 1	3	83.07	**88.07**	87.51
Citrate 2	3	84.45	**88.97**	88.66

Table 3.7: Log Pseudo Marginal Likelihood values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Largest values are shown in bold, indicating chosen model.

Metabolite	Actual Sites	1v2	1v3	2v3
Formate	1	0.03	0.05	1.74
Tris	1	0.04	0.05	1.29
Imidazole 1	1	0.03	0.06	1.70
Imidazole 2	1	0.04	0.08	1.72
Alanine	2	0.02	0.02	1.17
Creatinine 1	2	0.02	3E-3	0.14
Creatinine 2	2	0.03	0.04	1.57
Piperazine	2	3E-20	7E-20	2.26
Tartrate	2	0.03	0.04	1.39
Citrate 1	3	0.01	0.02	2.034
Citrate 2	3	0.02	0.02	1.29

Table 3.8: Bayes Factors calculated using approximate marginal likelihoods, and showing preference for either the 2 or 3 Site Models over the 1 Site Model.

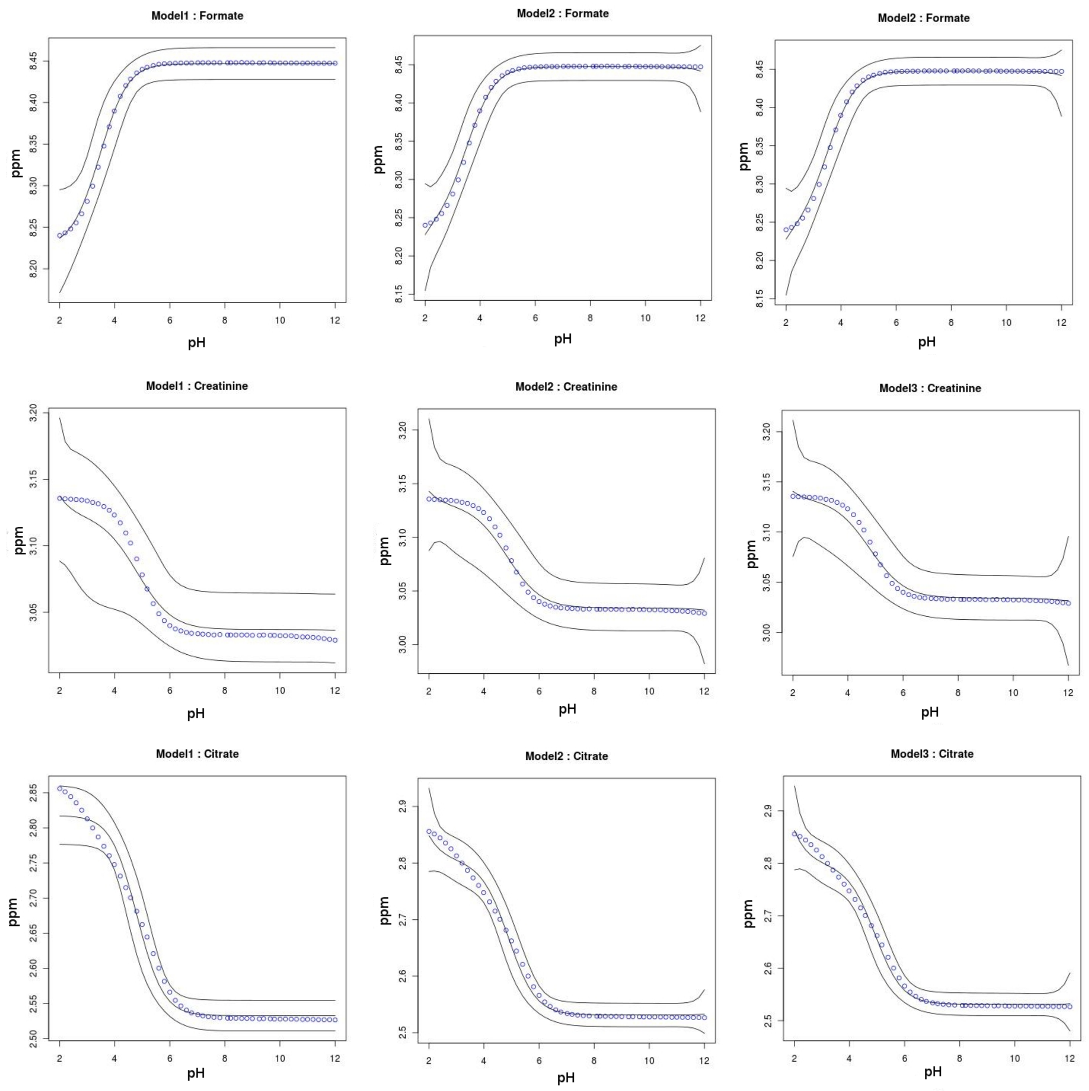


Figure 3.14: Polyprotic model fit for 1 Site molecule Formate (top), 2 Site molecule Creatinine (middle) and 3 Site molecule Citrate (bottom). Model1, Model2 and Model3 are polyprotic models with $q = 1, 2, 3$ sites respectively shown from left to right. The data is shown in blue, the fit and 95% credible intervals are indicated by black lines.

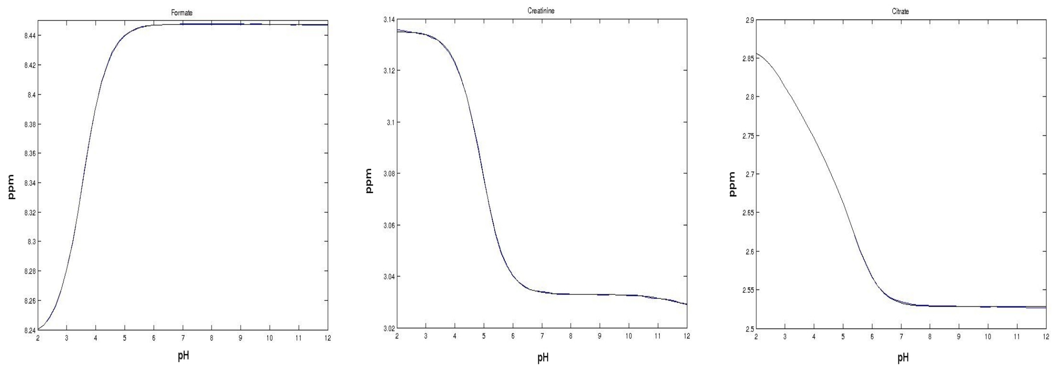


Figure 3.15: nlinfit model fit for 1 Site molecule Formate (left), 2 Site molecule Creatinine (middle) and 3 Site molecule Citrate (right) with data given in blue and the fit shown in black. The fit is so close as to be barely distinguishable by eye.

From Table 3.3, the AIC chooses the 3 Site Model regardless of the number of sites the metabolite has, except for Tris and Alanine where the 2 Site Model is selected. This suggests that the AIC does not penalise heavily enough on the number of parameters of the model. Similarly the BIC values shown in Table 3.4, indicate that the 3 Site Model is the 'best' fit, whilst again selecting the 2 Site Model for Tris and Alanine. It should be noted that AIC and BIC are not comparisons of the full Bayesian model but rather comparisons of the 'frequentist version' of the models using likelihood maximisation. Looking at Table 3.5, the DIC clearly has difficulty choosing between the three models which is reflected in the close values in the column of the table. The 2 Site Model is most consistently selected, indicating that the DIC may be penalising model complexity better than the AIC and BIC. However, looking at the estimate for the number of effective parameters of the model in Table 3.6 it is evident that the DIC method underestimates the number of parameters for the 2 and 3 Site Models suggesting the additional variables are correlated. A q Site Model has $2q + 2$ free parameters: q pKa values, $q + 1$ acid/base limits and the variance of the Normal likelihood. The differences between the LMPL values (Table 3.7) for each model are very small, showing that the LPML cannot really distinguish between models. The Bayes Factors are shown in Table 3.8. For all the metabolites there is some evidence to prefer either the 2 or 3 Site Model over the 1 Site Model and when comparing the 2 and 3 site we often prefer the model with fewer parameters.

Inability to correctly identify the model complexity is likely due to the sub-optimal fit provided by posterior parameter estimates symptomatic of the poor MCMC convergence in conjunction with the fact that there is not enough information in the data to estimate all the parameters unless we include further constraints.

3.4 Additive Constant

In order to improve the model fit, an additive constant, C, is introduced to μ to incorporate any additional unknown structured variance. So the q site model becomes:

$$\mathbf{y} \sim N(\mu(\mathbf{pH}, \mathbf{pKa}, \delta_{\mathbf{A}}, \sigma^2, q) \tag{3.26}$$

$$\mu_i = \frac{\delta_A + \Sigma_{m=1}^{q} \delta_{H_m A} 10^{(\Sigma_{j=q-m+1}^{q} pKa_j) - mpH_i}}{1 + \Sigma_{k=1}^{q} 10^{(\Sigma_{l=q-k+1}^{q} pKa_l) - kpH_i}} + C \tag{3.27}$$

$$C \sim N(0, t^2)$$

$$t^2 \sim IG(2, 0.5) \tag{3.28}$$

for $i = 1, .., n$ pH titrations and the rest of the model remains unchanged. The posterior pK$_a$ estimates are shown in Table 3.9.

Metabolite	Actual pKa	Posterior pKa Mean	Posterior pKa Variance
Formate	3.77	3.56	0.36
Tris	8.3	8.36	0.26
Imidazole 1	6.95	7.09	0.17
Imidazole 2	6.95	7.10	0.062
Alanine	2.35	3.99	4.48
	9.69	11.11	1.66
Creatinine 1	4.84	2.88	2.15
	9.2	8.77	4.17
Creatinine 2	4.84	2.09	2.06
	9.2	7.16	3.70
Piperazine	5.56	5.80	0.13
	9.83	10.10	0.19
Tartrate	2.98	1.54	1.15
	4.34	5.25	3.92
Citrate 1	3.09	1.11	1.29
	4.75	3.94	3.16
	5.41	8.94	4.31
Citrate 2	3.09	1.54	1.15
	4.75	5.25	3.92
	5.41	9.23	4.13

Table 3.9: Actual pK$_a$ values vs Posterior Mean pK$_a$ Estimates and Variance for model with additive constant.

3.4.1 Model Selection

Metabolite	Actual Sites	1 Site Model	2 Site Model	3 Site Model
Formate	1	-669.8	-686.0	**-722.9**
Tris	1	-447.6	**-608.0**	-607.85
Imidazole 1	1	-420.7	-561.6	**-616.5**
Imidazole 2	1	-332.7	**-486.0**	-482.0
Alanine	2	-273.3	-530.1	**-555.3**
Creatinine 1	2	-564.9	-664.4	**-666.2**
Creatinine 2	2	-559.3	-561.7	**-601.0**
Piperazine	2	-76.4	-388.8	**-480.2**
Tartrate	2	-380.3	-652.7	**-660.1**
Citrate 1	3	-245.3	-442.1	**-736.0**
Citrate 2	3	-267.5	-466.6	**-588.2**

Table 3.10: Akaike Information Criterion values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Smallest values are shown in bold, indicating chosen model.

Metabolite	Actual Sites	1 Site Model	2 Site Model	3 Site Model
Tris	1	-437.9	**-594.4**	-590.4
Imidazole 1	1	-411.0	-548.1	**-599.2**
Imidazole 2	1	-323.0	**-472.4**	-464.6
Formate	1	-660.1	-672.4	**-705.5**
Alanine	2	-263.6	-516.6	**-537.9**
Creatinine 1	2	-555.3	**-650.8**	-648.8
Creatinine 2	2	-549.7	-548.2	**-583.6**
Piperazine	2	-66.8	-375.2	**-462.8**
Tartrate	2	-370.9	-639.6	**-643.2**
Citrate 1	3	-235.6	-428.5	**-718.6**
Citrate 2	3	-257.8	-453.1	**-570.8**

Table 3.11: Bayesian Information Criterion values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Smallest values are shown in bold, indicating chosen model.

Metabolite	Actual Sites	1 Site Model	2 Site Model	3 Site Model
Tris	1	**91.4**	90.4	89.8
Imidazole 1	1	**91.4**	90.6	90.3
Imidazole 2	1	**90.7**	90.1	89.6
Formate	1	**91.2**	90.6	89.5
Alanine	2	88.6	**89.9**	89.5
Creatinine 1	2	89.9	**90.2**	88.1
Creatinine 2	2	**91.5**	90.8	90.2
Piperazine	2	38.6	**89.1**	88.4
Tartrate	2	91.2	**83.4**	83.2
Citrate 1	3	86.7	**87.83**	87.75
Citrate 2	3	88.4	**88.5**	87.0

Table 3.12: Log Pseudo Marginal Likelihood values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Largest values are shown in bold, indicating chosen model.

Metabolite	Actual Sites	1v2	1v3	2v3
Tris	1	2.45	5.04	2.05
Imidazole 1	1	1.98	2.65	1.34
Imidazole 2	1	2.91	6.13	2.11
Formate	1	1.47	2.84	1.69
Alanine	2	0.40	0.61	1.54
Creatinine 1	2	1.50	2.70	1.80
Creatinine 2	2	2.16	3.94	1.82
Piperazine	2	2.81E-20	6.84E-20	2.43
Tartrate	2	0.029	0.031	1.06
Citrate 1	3	0.51	0.53	1.03
Citrate 2	3	2.12	2.20	1.03

Table 3.13: Bayes Factors calculated using approximate marginal likelihoods, and showing preference for either the 2 or 3 Site Model.

Metabolite	Actual Sites	1 Site Model	2 Site Model	3 Site Model
Tris	1	**-179.3**	-177.3	-175.5
Imidazole 1	1	**-179.3**	-177.0	-175.8
Imidazole 2	1	**-179.3**	-176.8	-175.6
Formate	1	**-179.4**	-175.7	-174.1
Alanine	2	-174	**-174.1**	-173.5
Creatinine 1	2	**-176.3**	-174.5	-174.6
Creatinine 2	2	**-179.6**	-177.5	-176
Piperazine	2	-74.75	**-172.8**	-170.7
Tartrate	2	**-164.5**	-163.1	-162.3
Citrate 1	3	-169.7	**-171.5**	-170.5
Citrate 2	3	-172.9	**-173.2**	-172.2

Table 3.14: Deviance Information Criterion values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Smallest values are shown in bold, indicating chosen model.

Metabolite	Actual No. of Params.	1 Site Model p_D	2 Site Model p_D	3 Site Model p_D
Tris	6	4.24	5.02	5.43
Imidazole 1	6	4.22	5.07	5.46
Imidazole 2	6	4.20	5.01	5.57
Formate	6	4.19	4.86	5.40
Alanine	8	4.18	5.53	5.88
Creatinine 1	8	5.19	5.73	5.71
Creatinine 2	8	4.14	4.86	5.36
Piperazine	8	5.13	6.28	7.00
Tartrate	8	4.42	4.92	5.11
Citrate 1	10	4.75	5.83	6.37
Citrate 2	10	4.63	5.58	6.03

Table 3.15: Actual number of free parameters for appropriate model and p_D values (estimated number of effective parameters) for each model.

Once again, AIC (Table 3.10), BIC (Table 3.11), and Bayes Factor (Table 3.13) struggle to choose between the 2 Site and 3 Site Models, although the Bayes Factor seems to perform a little better for the additive model, whilst LPML (Table 3.12) marginally prefers the 2 Site Model. DIC (Table 3.14) now consistently selects the less complex models (1 Site Model for all except Citrate 1 and 2 and Piperazine), and Table 3.15 shows the number of parameters are still underestimated. Looking across the two sets of results we see that the 1 Site Additive Model is generally preferred to the 2 Site Model (without additive constant) suggesting the additive constant is compensating somewhat for the fact that the fit is not as good as perhaps it could be.

3.5 Random Effects Model

Next, strength is borrowed across chemical shifts of the same metabolite to estimate the pK$_a$ in an attempt to improve the accuracy of the pK$_a$ estimates and general model fit [98]. The pK$_a$ value should be the same for all chemical shifts of the same metabolite so a random effects term is used for the pK$_a$ mean, giving us the following model for a metabolite with S chemical shifts:

$$\begin{align}
\mathbf{y}_s &\sim N(\mu_s(\mathbf{pH}, \mathbf{pKa_s}, \delta_{\mathbf{A_s}}, \sigma_s^2, q)) \tag{3.29}\\
\mu_{s,i} &= \frac{\delta_{A_s} + \Sigma_{m=1}^{q} \delta_{H_{s,m}A} 10^{(\Sigma_{j=q-m+1}^{q} pKa_{s,j}) - mpH_i}}{1 + \Sigma_{k=1}^{q} 10^{(\Sigma_{l=q-k+1}^{q} pKa_{s,l}) - kpH_i}} + C_s \tag{3.30}\\
& \tag{3.31}
\end{align}$$

for $s = 1, .., S$ chemical shifts consisting of measurements at $i = 1, .., n$ different pH values, where $\mu_s = (\mu_{s,1}, ..., \mu_{s,n})$ is the mean vector for the s-th chemical shift. $\mathbf{pKa_s} = (pKa_{s,1}, ..., pKa_{s,q})$ denotes the q pK$_a$ values and $\delta_{\mathbf{A_s}} = (\delta_{A_s}, \delta_{H_{s,1}A}, ..., \delta_{H_{s,q}A})$ gives the $q+1$ chemical shift limits for the s-th chemical shift of the metabolite. Constructing prior distributions as before, gives:

$$\begin{align}
\sigma_s^2 &\sim IG(\alpha/2, \beta/2) \tag{3.32}\\
pKa_{s,j} &\sim N(P_j, 1) \notag\\
P_j &\sim U(0, 14) \notag\\
\delta_{A_s} &\sim U(0, 14) \notag\\
\delta_{H_{s,j}A} &\sim U(0, 14) \notag\\
C_s &\sim N(0, t_s^2) \notag\\
t_s^2 \sim IG(2, 0.5) j &= 1, .., q \notag\\
s &= 1, .., S \notag
\end{align}$$

Three metabolites with two chemical shifts each were chosen for further investigation: Imidazole, Creatinine and Citrate with 1, 2 and 3 sites respectively. The posterior pK$_a$ estimates are given in Table 3.16. As is clear from the table, there is no improvement in the estimates.

Metabolite	Actual pKa	Posterior pKa Mean	Posterior pKa Variance
Imidazole	6.95	7.094842	0.5391012
Creatinine	4.84	2.370332	4.385893
	9.2	8.091135	20.69006
Citrate	3.09	0.5457603	0.2978543
	4.75	1.811553	2.346749
	5.41	6.09283	10.55541

Table 3.16: Actual pKa values vs Posterior Mean pKa Estimates and Posterior pKa Variance for Random Effects Model.

3.5.1 Model Selection

As can be seen in Table 3.17, the DIC selects the 1 Site Model for all three metabolites. Looking at Table 3.18 it is clear that the number of free parameters is still not well estimated.

Metabolite	1 Site Model	2 Site Model	3 Site Model
Imidazole	**-386.3**	-382.9	-381.6
Creatinine	**-385.9**	-381.5	-380.4
Citrate	**-382.8**	-380.6	-379.6

Table 3.17: Deviance Information Criterion values for 1, 2 and 3 Site Models fitted to all the metabolites, alongside actual number of sites of each molecule. Smallest values are shown in bold, indicating chosen model.

Metabolite	Actual No. of Params.	1 Site Model p_D	2 Site Model p_D	3 Site Model p_D
Imidazole	13	9.258	10.46	10.86
Creatinine	18	9.297	10.64	10.93
Citrate	23	9.933	11.06	11.38

Table 3.18: Actual number of free parameters for appropriate model and p_D values (estimated number of effective parameters) for each model.

3.6 Conclusion

Whilst the polyprotic model provides a good fit to the titration curves and fair estimates of pK$_a$ values, it is clear that several different model selection criteria all have difficulty in correctly predicting the number of sites the metabolite has. The introduction of an additive constant did little to improve the accuracy of parameter estimates, and it still proved difficult to select the correct model. The random effects model also showed no improvement in parameter estimation or model selection. These results indicate that predicting the number of sites of a molecule from its titration curve is a difficult problem needing much further investigation.

Chapter 4

LC-MS Processing

4.1 Background

In this chapter I review some of the challenges faced and methods used in LC-MS data pre-processing and also outline some new strategies. Since LC-MS is a technique popular within proteomics as well as metabolomics, some pre-processing techniques used across both fields are examined. Given the nature of vast raw LC-MS datasets, most pre-processing algorithms first chop the data into mass-to-charge ratio (m/z) bins and interpolate in the retention time (Rt) dimension over equally spaced nodes. If more than one Rt data point is swallowed by a single (m/z, Rt) bin, then either the largest of the intensities or the sum is used. The resulting dataset can therefore be processed as a matrix of intensities. A by-product of binning is that the data can be arranged into a series of 'Extracted Ion Chromatograms' (EICs). XCMS [46], a widely used pre-processing package, scales these chromatograms by the largest peak resulting in a set of 'Extracted-Ion Base Peak Chromatogram' (EIBPCs).

The following conceptual model of the chromatographic signal is common: $y_i(t) = b(t) + N \times s(t) + \epsilon(t)$ for $i = 1, ..., n$ where $b(t)$ is the baseline, $s(t)$ is the true signal, N is a normalisation factor and $\epsilon(t)$ is noise-primarily electronic resulting from the detector and dependent on t (Retention Time) and n is the number of chromatograms [108]. This approach means that pre-processing is performed on each EIC (or sometimes mass spectrum if the m/z dimension is processed first) individually rather than borrowing strength across both dimensions, although there are several algorithms akin to 2-dimensional image processing that attempt to buck this trend [3, 109, 110].

It is generally accepted that essential pre-processing steps are Retention Time correction (or alignment), peak detection, noise reduction or background subtraction, peak quantification, pattern detection and matching, which can be enhanced by combining replicate datasets to improve SNR and also systematically characterising sources of variability [50]. There is some debate as to the optimum order to perform these processes and also considerable variation of methods used within each of them. For example, there is the question of whether to perform Retention Time alignment before or after peak detection since peaks positions could aid alignment, whether to use the continuous signal or peak positions, whether or not the process should be dependent on signal amplitude and if samples should be aligned to some pre-defined alignment 'template' or to the mean of all the samples. Alignment across samples can take a variety or forms, e.g. correlation optimised warping, vectorised peaks, (semi) supervised alignment using non-linear regression models, hidden markov models, statistical alignment or clustering [50, 111, 112, 113]. One of the most widely-used method is Correlated Optimised Warping (COW) which performs a linear warp within a window in order to optimise overlap whilst maintaining matched boundaries.

Filtering and background subtraction is either performed by subtracting a fitted, additive baseline model (e.g. splines) or using digital filters to smooth and enhance the MS signal (e.g. moving average, Savitsky-Golay, matched filtering etc.). These can be performed on one or both of the dimensions (Rt or m/z) simultaneously, alternately or iteratively. There exist myriad algorithms utilising different combinations of peak models and criteria, including searching for signal maxima that coincide in both the m/z and Rt directions (vectorised peak detection) [114], intensity thresholding, examining the local neighbourhood of a candidate peak, using wavelet decomposition to investigate scale-specific peaks, or curve resolution to extract major spectral components [115, 116]. Often, signal filtering and background subtraction are performed implicitly when detecting peaks.

4.1.1 Wavelet Smoothing

Wavelets have become a popular choice of method for de-noising a variety of signals and are becoming widely used within metabolomics. Wavelets are mathematical functions resembling 'small waves' and are particularly useful in image compression, where the resulting decomposition is often sparse. Wavelet decomposition is a multi-scale method, meaning the decomposition consists of coefficients describing the data object at a range of dilations and positions of the wavelet function used. The wavelet function is chosen to reflect the shape of wave characteristic of the data (see Figure 4.1 for examples).

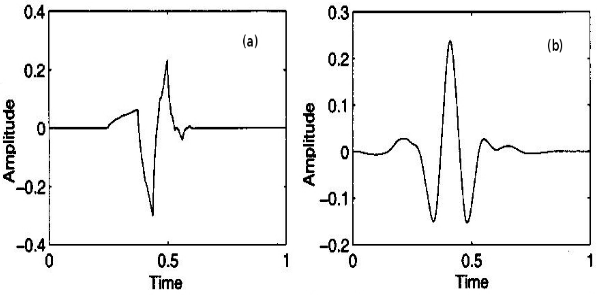


Figure 4.1: Examples of two commonly used wavelet functions (a) Daubechies 4, (b) Symmlet 8 [4]

Let $\psi_{a,b}(t), a \in \mathbb{R}^+, b \in \mathbb{R}$, denote a family of wavelets consisting of translations and dilations of a 'mother wavelet', $\psi(t) \in L_2(\mathbb{R})$ [117]:

$$\psi_{a,b}(t) = \mid a \mid^{-1/2} \psi(\frac{t-b}{a}) \tag{4.1}$$

For a function of finite energy, f, the Continuous Wavelet Transform is defined by C_f [118]:

$$C_f(a,b) = \int_{\mathbb{R}} f(t)\bar{\psi}_{a,b}(t)dt \tag{4.2}$$

where $\bar{\psi}$ represents complex conjugation of ψ. The family of wavelets must satisfy the admissibility condition, which specifies $\psi(x)$ should be oscillatory and localised in time and frequency in addition to integrating to zero:

$$\int_{-\infty}^{\infty} \frac{\mid \Psi(\omega) \mid^2}{\mid \omega \mid} d\omega < \infty \tag{4.3}$$

where $\Psi(\omega)$ is the Fourier transform of $\psi(t)$.

In order to reduce the redundancy of the transform, the values of a and b can be discretised whilst retaining the invertibility of the transform. 'Critical sampling', or the finest scale possible, produces a basis on L_2:

$$\begin{aligned} \{\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k), j, k \in Z\} \\ a = 2^{-j} \\ b = 2^{-j}k \\ j, k \in Z \end{aligned} \tag{4.4}$$

This convention enables implementation of the Discrete Wavelet Transform, a fast decomposition algorithm (see [4] for details). Decimation, which enables coarse wavelet coefficients to be described as a convolution of finer scale coefficients can reduce redundancy further; however, the Undecimated Discrete Wavelet Transform (UDWT) is often preferred since it is shift-invariant.

The wavelet coefficients of fine scales correspond to high frequency features in the original signal, so these can be thresholded before performing the inverse transform. There are a number of threshold methods available, each with specific advantages and disadvantages. The main methods for applying a threshold λ are 'soft' (Equation 4.5) and 'hard' (Equation 4.6) thresholding of a signal $y(t)$:

$$y_{soft}(t) = \begin{cases} sign(y(t)) \cdot (|y(t)| - \lambda) & \text{if } |y(t)| > \lambda \\ 0, & \text{if } |y(t)| < \lambda \end{cases} \tag{4.5}$$

$$y_{hard}(t) = \begin{cases} y(t) & \text{if } |y(t)| > \lambda \\ 0 & \text{if } |y(t)| < \lambda \end{cases} \tag{4.6}$$

where the function *sign* determines whether a value is positive or negative. When the wavelet reconstruction is performed, the noise should be removed and the signal visibly cleaner.

4.1.2 Peak Detection

Here I discuss some of the most commonly used techniques for peak detection.

Vectorised Peak Detection

Perhaps the most intuitive way to distinguish analyte peaks from noise, is to look for peaks that are local maxima in both the retention time dimension and the m/z dimension. This approach is formalised by Hastings et al. [119], but many other peak detection algorithms use this criterion in conjunction with others, e.g. MZmine also imposes specified peak widths [120]. An advantage of this simple technique are that points lying in ion chromatograms corrupted by solvent clusters or column bleeds are only identified if they exceed a threshold defined by the level of noise in that particular chromatogram, not adjacent ones.

Isotopic Pattern Matching

Whilst generally concerned with larger molecules (more typically within proteomics than metabolomics), an interesting area of investigation is de-isotoping or charge deconvolution. This approach uses the fact that isotopic peaks often appear in LC-MS data resulting from heavy isotopic variants of carbon and other atoms, forming a predictable pattern [50]. Based on the mass differential, one can deduce the charge state of multiple charged ions and then compare with reference tables. However, this is difficult to perform on complex mixtures. There are several peak detection algorithms utilising isotopic pattern matching, for example VIPER [121, 122], MSInspect [123], PEPLIST [124]. VIPER uses an isotopic pattern matching routine, THRASH, in conjunction with a database of previously identified molecule positions [121, 122]. THRASH, like most isotopic pattern matching algorithms, works by comparing an 'expected isotope pattern' to this observed isotopic cluster, and producing a 'fitting score'. Methods for generating the theoretical isotopic distribution vary, from using a Poisson distribution [123] to a model amino acid 'averagine' (an 'average' amino acid calculated from occurences of amino acids in a protein database) in proteomics applications [124].

Image Detection

Another area of research being exploited for methods applicable to LC-MS peak detection, is mathematical morphology and image detection theory. Mathematical morphology is concerned with the characterisation of geometric structures using set theory, lattice theory, topology and random functions [125, 126, 127, 128]. Packages adopting these techniques, such as edge-detection and segmentation functions, include MeDDL [109], DoGEX [110] and MapQuant [3]. Given the 2D nature of LC-MS data, it can be useful to visualise a sample as a heat map image, where intensity defines colour. Indeed Du et al. (2007) have developed a supervised package, LC-MS-2D [129], in order to take advantage of this fact.

Wavelets

There are several peak detection packages using wavelets in conjunction with other peak-picking criteria or thresholding, or directly for peak picking themselves [123, 124, 130, 131, 132, 133]. For example, PrepMS [134] is a pre-processing package that de-noises MS spectra using an undecimated wavelet transform as described by [108] and detects features using the mean spectrum as outlined by [135]. Also developed by Morris, is Pinnacle: a wavelet-based package for detecting and quantifying protein spots. Gel-based electrophoresis used in proteomics creates datasets that share some of the 2-D complexity of metabolomic LC-MS data. Morris de-noises the average gel using wavelet shrinkage and hard thresholding. All peaks are detected from the reconstructed signal and those within a given proximity of a higher intensity are then combined

[136].

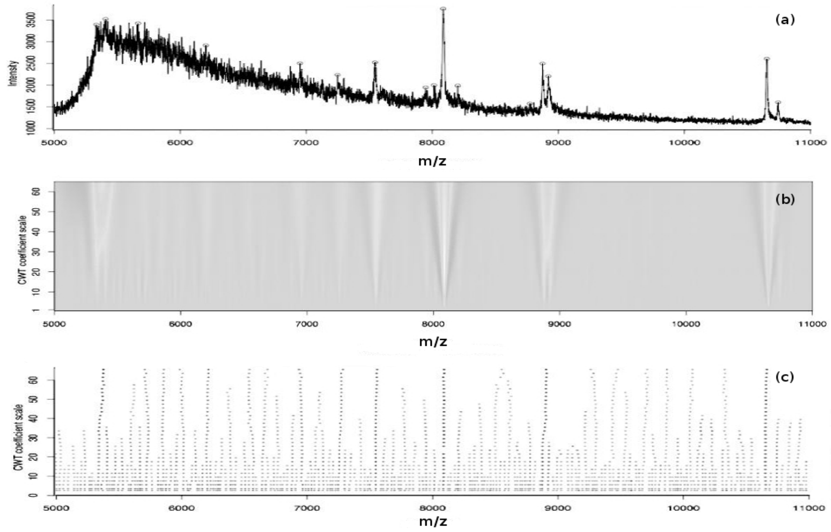


Figure 4.2: (a) Identified Peaks in m/z domain with SNR >3, (b) CWT coefficients of the signal at different scales, (c) 'Ridge Lines' constructed from joining the location of local maxima at different scales and indicate where 'important' features persist over multiple scales.[5]

By comparison, Du, W. et al. (2006) employs a continuous wavelet transform using the Mexican Hat wavelet (Gaussian second-derivative) and identifies peaks by tracking the ‘ridge lines’ of a colour-map representing wavelet coefficients at different scales (see Figure 4.2)[5]. Identified peaks (ridge lines) can be further thresholded using the scale, SNR or ridge length.

Kalman Tracking

An example of digital filtering is the Kalman Tracking algorithm used by Aberg [137]. The aim is to circumvent the problem of mass binning by tracking features in the two dimensional plane. A centroid mode mass spectrum from one scan can be viewed as a set of points on a 2D plane with m/z as x-axis, y-axis as intensity. The data will consist of structured signal which should behave predictably, and random noise. Given this data structure it is possible to borrow algorithms from tracking signals of objects, such as aircraft, in radar data. In constructing the Kalman Filter, there are three considerations:

- Movement of the object, i.e. process uncertainty.
- Measurement uncertainty - variation resulting from imprecise measurement technique.

- Noise - Random signal interference.

In the short term the object's trajectory is expected to be smooth, with long term changes in direction so search for an optimal filter for tracking an object with constant velocity in the presence of random velocity changes between measurements. One has no knowledge before measurements of how the object will manoeuvre. The structure of consecutive centroid mode mass spectra (scans) can be related to radar scans where m/z, and intensity are linked to azimuth (angle of trajectory) and range respectively.

Functional Mixed Models

Functional mixed models can identify differentially expressed regions possibly missed by peak detection and can automatically adjust for nuisance factors [108, 138]. In [139, 140] Morris sets out a novel way of investigating LC-MS data within a Bayesian framework. Functional mixed modelling avoids reliance on peak detection methods whilst allowing flexibility and spatially adaptive regularisation as well as being able to simultaneously model multiple effects. Some pre-processing, baseline correction, normalisation, de-noising and possible transformation, may be carried out prior to model fitting. Morris' model is characterised by the following equation:

$$y_i(t) = \Sigma_{j=1}^{p} X_{ij} V_j(t) + \Sigma_{k=1}^{m} Z_{ik} U_k(t) + E_i(t) \tag{4.7}$$

$$\mathbf{U}(t) \sim N(0, Q) \tag{4.8}$$

$$\mathbf{E}(t) \sim N(0, S) \tag{4.9}$$

where y_i, for $i = 1, .., N$, represent N functions (in this case MALDI-TOF spectra) defined over the time axis t. (X_{ij}) is the matrix of covariates, V_j for $j = 1, ..., p$ the vector of p functional fixed effects, (Z_{ik}) is the design matrix for the m functional random effects $\mathbf{U}(t) = (U_1, ..., U_m)$ with covariance Q and $\mathbf{E}(t) = (E_1, ..., E_n)$ with covariance S is the residual error. A discrete matrix version of the model can be described and is fitted using a basis function approach, using wavelet basis functions since they characterise the MS peaks well.

The discrete wavelet transform (DWT) is applied to each of the N spectra, projecting the observed data into the wavelet-spanned space. MCMC samples are obtained to the wavelet-space version of the model which are then projected back to the data-space using the IDWT. Bayesian inference can then be applied to these samples, for example posterior probabilities of being a feature of interest can be computed for each peak, and a threshold of significance determined.

Diffuse proper priors are specified for variance components and adaptively regularised representations of the fixed effect functions V_j are obtained by utilising a multiple-shrinkage prior on the wavelet coefficients for the fixed effects.

Combining Criteria

A number of LC-MS pre-processing packages, for example MZmine [120, 141], Yasui et al. [142], PEPLIST [124], MSInspect [123], integrate two or more of the approaches mentioned above, often combining peak criteria to give an overall ‘fitting score’. Some of the peak characteristics may also be thresholded, for example peak intensity, peak length or shape (by fitting a function such as a Gaussian) in one or both dimensions. One such package is MEND [143] which incorporates matched filtering, vectorised peak detection and isotopic pattern matching. A fitting score for each of the three characteristics is calculated and the scores are thresholded.

Full processing suites may also give access to a number of different techniques for each of the pre-processing steps required (including peak detection) that can be performed separately or together, for example MZmine. The more recent overhaul of MZmine resulted in a modular package containing multiple methods for smoothing, peak detection, peak alignment, peak identification, normalisation, visualisation and other statistical analysis [141]. There are multiple ‘plug-in’ algorithms for peak detection including local maxima searching, a method based on the mexican hat wavelet transform, SNR thresholding, filtration, LC peak length thresholding and intensity thresholding.

Conclusion

The methods described in this section and Table 4.1 are by no means an exhaustive list, there has been much recent research into other multivariate approaches [115], second order methods [144] and MCMC algorithms [145]; but Wavelets, Digital Filters, Isotopic Pattern Matching and Image Processing techniques remain the more standard approaches.

Within proteomics applications [146] finds that Wavelet-based methods perform better than other filters combined with more heuristic criteria for 1D MS peak detection. [147] also promotes wavelet smoothing as characterising peaks shapes well since by smoothing over multiple scales one avoids the problem of more traditional filters that use a fixed size window. By contrast, [47] evaluates filters for smoothing in the chromatographic domain and finds that the filter performance (measured by SNR) is very much dependant on data characteristics such as noise level and type as well as peak size, shape and integrity. [45] reports that the best performing algorithms for LC-MS are those that utilise both the m/z and Rt domains to locate features and thus more fully exploit the 2-D nature of the data.

However, incorporating more prior knowledge about the data structure does not necessarily result in a more effective algorithm. Packages based on simple intensity thresholding, for example msInspect, have been shown to outperform those using more descriptive peak picking criteria, such as MZmine. [41] shows how combining the intensity thresholding with other peak picking criteria such as thresholding LC peak length or isotope pattern matching did nothing to improve the performance. More complex LC peak filtering also underperformed compared to the peak intensity-based method since the peaks are often poorly shaped and do not resemble the traditional chromatographic peak models, such as Gaussian.

There exist many reviews documenting the abundant peak detection algorithms and methods on offer but few articles that rigorously compare or critique their performance alongside each other, particularly within the context of metabolomics data. Many of the algorithms were initially designed to fulfil the particular analytical needs of a given lab (i.e. specific LC-MS or MS configuration, data format, bio-fluid/organism type, context of experiment and application - perhaps proteomics rather than metabolomics etc.) and are later generalised or else are an exploration of a particular statistical/modelling technique but not presented as part of a full pre-processing tool. For those that are available as an end-to-end processing suite applicable across different (LC-)MS setups, comparison cannot necessarily be performed fairly given the variability of algorithms employed for the other pre-processing steps such as retention time alignment, de-noising etc. Peak detection performance may also depend on the order in which the steps are executed and will often be inextricably coupled to one or more of these other steps so cannot be compared in isolation. Both [41] and [50] mention the problem that many peak detection algorithms are of a relatively ad-hoc nature and as such no single package seems to offer a universally robust solution or one that does not require substantial parameter tuning.

Generally speaking, choice of peak detection method will depend on the instrument used (most manufacturers also provide their own closed processing software specific to the machine), the output data format (different software packages handle different file formats), characteristics of the data (if background noise a particular problem for the experiment etc. it may need an algorithm employing more aggressive smoothing techniques before peak detection than for another experiment) and also personal preference of the person or lab where analysis is being performed (the individual's expertise may be within a particular statistical environment such as R/MATLAB etc. or a tool may have been developed within the lab). Researchers may also choose to run their data through multiple software to prevent feature loss that may occur as a result of using a single algorithm, on the assumption that different approaches may be complimentary and perhaps pick up aspects of the data that a different algorithm missed.

LC-MS technology is advancing at a rapid rate with instruments producing larger and more complex datasets as a result of capability for much greater sensitivity and resolution. This in itself poses a challenge as software and processing techniques must either evolve quickly to keep up or be robust enough to cope with changing data characteristics.

Given the huge variability of a LC-MS dataset due to the multiple factors described, one could argue that the more powerful techniques are those that are robust to changing noise levels, peak size and shape and/or can self-learn the characteristics of the data it is processing. However, there is useful information to be derived from knowing the configuration of the instrument used and characteristics of the biofluid so perhaps more complete feature detection can be achieved when the algorithm is finely tuned or even specifically designed to capture these nuances for a particular experimental set-up, or at least until machine learning algorithms can truly catch up with the expertise of the researcher in recognising veritable features.

Package	Platform	Data	Methods	Ref.
apLC-MS	LC-MS	Metabolomics	Adaptive Binning	[66]
centWave	LC-MS	Both	Wavelet Denoising	[130]
DoGEX	LC-MS	Metabolomics	Wavelet Denoising Image Detection	[110]
LC-MS-2D	LC-MS	Proteomics	Isotope Pattern Matching	[129]
MapQuand	LC-MS	Proteomics	Matched Filtration Mathematical Morphology	[3]
MeDDL	LC-MS	Metabolomics	Image Detection	[109]
MEND	LC-MS	Proteomics	Matched Filtration Isotope Pattern Matching Vectorised Peak Detection	[143]
msInspect	LC-MS	Proteomics	Local Maxima Isotope Pattern Matching	[123]
MZmine	LC-MS	Both	Vectorised Peak Detection Image Detection	[120]
PepList	LC-MS	Proteomics	Wavelet Denoising Isotope Pattern Matching	[124]
Pinnacle	2DE	Proteomics	Wavelet Denoising Mean Spectrum	[140] [135]
PrepMS	TOF-MS	Proteomics	Wavelet Denoising Local Maxima	[134]
TracMass	GC/LC-MS	Metabolomics	Kalman Tracking Filter	[137]
VIPER	LC-MS	Proteomics	(Rt, m/z) Tag Database Isotope Pattern Matching	[122] [121]
WaveletQuant	LC-MS	Proteomics	Wavelet Denoising	[132]
waveSpec	MS	Both	Wavelet Denoising Kernel Density Estimation	[148]
wfmm	Functional Data	Both	Wavelet-Based Functional Mixed Models	[140] [139]
XCMS	LC-MS	Metabolomics	Matched Filtration or Wavelet Denoising	[46]
Befekadu et. al	LC-MS	Proteomics	Functional Mixed Models	[138]
Du, et. al	SELDI TOF MS	Proteomics	Wavelet Domain Local Maxima	[5]
Hastings et. al	LC-MS	Both	Vectorised Peak Detection	[119]
Stolt et. al	LC-MS	Metabolomics	Second Order Method	[144]
Wang et. al	SELDI-MS	Proteomics	Functional Mixed Models	[145]

Table 4.1: Overview of existing pre-processing algorithms for (LC-)MS data.

4.1.3 Data Binning

As technological advances allow us capability for higher resolution LC-MS data the problem of data binning becomes ever more apparent. High resolution data files are extremely large and unwieldy from a data analysis point of view, often containing a large proportion of chemical and instrumental noise. Myriad smoothing and feature extraction algorithms have been developed and this remains a thriving area of research. However, less literature is devoted to the problem of data binning. It is an optimisation problem that is often overlooked in favour of uniformly spaced bins of a size specified by the user. Peaks can drift in and out of bins if the bin size is too small (see Figure 4.3), if a cut-off point intersects with a peak or, equally undesirable, multiple peaks can be squashed into a single bin if the bin size is too large and information is lost. Smaller bins create extra computations, but they are more likely to capture fewer peaks per bin. Smith et al. [46] attempt to address this problem in XCMS by constructing overlapping bins and filtering out duplicate peaks under supervision.

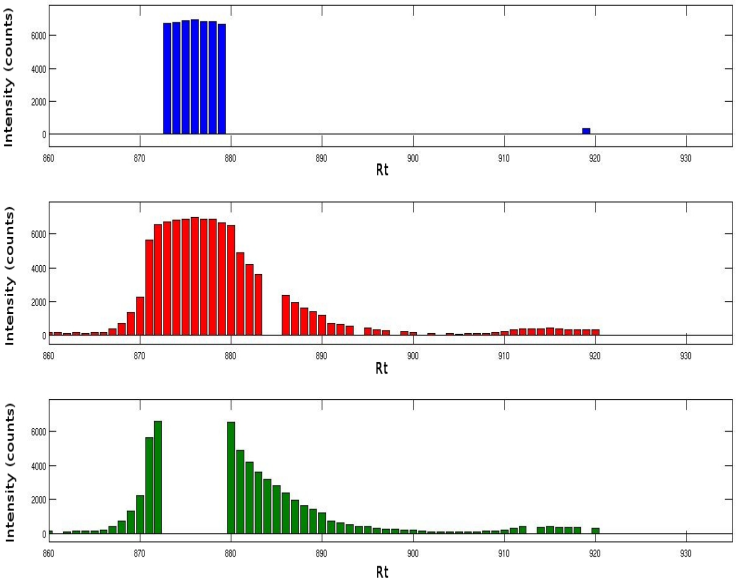


Figure 4.3: A chromatographic peak split across three bins due to drift in m/z.

Another package that attempts to tackle this is apLC-MS [66]. apLC-MS orders the data-points by m/z value and then constructs a vector containing the differences between consecutive points. A mixture model describing the differences in m/z value between and within peaks is constructed to calculate an m/z threshold and the data grouped accordingly (described in further detail later in the chapter). Similarly to LC-MS, drift is also a problem within NMR spectra and binning is often used in order to evaluate spectral density. Although this problem is 1-Dimensional, Davis et.al [149] take an interesting approach using wavelets which is worth noting. Similarly to the method used by Du et.al in centWave [5] for detecting maxima in chromatograms, Davis uses a wavelet transform (undecimated, to avoid filter artifacts) but detects minima in the wavelet domain instead. The aim is to identify minima that indicate the beginning and end of each peak. In contrast to Du, Davis performs the transform at a single level rather than linking minima across levels, with the appropriate level depending upon the data resolution and typical peak width. In Section 4.3 I build on these methods to develop a novel binning strategy as part of a new pre-processing algorithm, H-MS, but first a Bayesian model for peak detection is explored.

4.2 A Bayesian Partition Model for Peak Detection

We propose a Bayesian strategy to analyse LC-MS peak distributions using a 'Bayesian Partition Model' (BPM) also known as a 'Change Point Model'. The goal is to develop an LC-MS algorithm that identifies co-eluting peaks that may belong to the same molecular species. To this end, we will need to de-noise the data as well as identify and describe metabolite peaks as accurately as possible.

BPMs have been employed increasingly in a variety of settings, from cosmology to economics [150, 151, 152]. The model assumes that observations within a segment of a random partition of the series follow the same distribution [153]. They were pioneered by Barry and Hartigan in the 1990s and have since gained popularity in many fields [154, 155, 156, 157, 158, 159, 160].

Initially working on individual chromatograms, where the n observations are denoted by $\mathbf{y} = (y_1, ..., y_n)$, a BPM is developed to characterise changes in signal level. This way, if each segment follows a normal distribution with parameters dependent on the partition, a peak should be 'detected' as a segment with a mean level significantly different from zero. The change-points $\tau = (\tau_1, ..., \tau_k)$, specify the data-points at which a new segment of the partition begins. Each possible number of change-points, k, will have multiple change-point configurations. It is the posterior distribution of the change-point locations that will tell us the most likely points at which the mean signal level changes significantly thus identifying possible peak locations. We denote the signal segments by R_i, defined by Equation 4.10:

$$\begin{aligned} R_i = \{y_j : j = \tau_{i-1}+1,..,\tau_i\} \text{ for } i = 2,..,k \\ \text{and } R_1 = \{y_j : j = 1,..,\tau_1\} \\ R_{k+1} = \{y_j : j = \tau_{n-1}+1,..,n\} \end{aligned} \tag{4.10}$$

Let $\tau^{(k)}$ be the set of all possible partitions over the locations $\{2,..,n-1\}$ of size $k \in \{0,..,n-2\}$ change points. All possible partitions for a given k are equally likely so we assign a uniform prior to $\tau^{(k)}$. We also assign a truncated geometric prior to k:

$$\begin{aligned} p(\tau^{(k)}, k) = p(\tau^{(k)} \mid k)p(k) \\ p(\tau^{(k)}|k) = \binom{n-2}{k}^{-1} \\ p(k) = (1-q)^k q\mathrm{I}(0, n-2) \end{aligned} \tag{4.11}$$

where q is a constant parameter. Given this partition, the likelihood is simply the product of the likelihood for each of the $k+1$ segments:

$$p(\mathbf{y} \mid \tau^{(k)}, k) = \prod_{i=1}^{k+1} p(\mathbf{y}_{R_i} \mid \tau^{(k)}, k) \text{ for } \mathbf{y}_{R_i} = \{y_i \in R_i\} \tag{4.12}$$

Likelihood

We assign $y_{R_i} = \{y_j \in R_i\}$ for $i = 1,..,k+1$ a multivariate normal distribution:

$$\mathbf{y}_{R_i} \mid \tau^{(k)}, k \sim N(\mathbf{x}_i' \mu_i, I\sigma_i^2) \text{ for } i = 1, ..., k+1 \tag{4.13}$$

where $\mathbf{y}_{R_i}$ is a vector of length n_i. $\mathbf{x}_i'$ is the 'design' matrix, a $1 \times n_i$ vector describing the shape of the data. μ_i and σ_i are scalar parameters, whilst I is the appropriate identity matrix. For computational simplicity, we will initially use a vector of 1's for the design matrix.

Conjugate Prior

We assign conjugate a prior to μ_i as follows:

$$\mu_i \sim N(\theta_i, p_i \sigma_i^2) \tag{4.14}$$

where the prior mean, θ_i, and precision, p_i, are constants. Again, using conjugacy we have the following prior for σ_i:

$$\sigma_i^2 \sim IG(\frac{d_i}{2}, \frac{s_i}{2}) \tag{4.15}$$

where $\frac{d_i}{2}$ is the shape parameter and $\frac{s_i}{2}$ is the scale parameter.

Marginal Likelihood

Using a useful theorem for normal-normal models, given in Appendix .3, in conjunction with another theorem for marginalising over μ we have:

$$\mathbf{y}_{R_i} \mid \sigma_i \sim N(\mathbf{x}_i'\theta_i, \sigma_i^2(p_i\mathbf{x}_i'\mathbf{x}_i + I)) \tag{4.16}$$

$$i = 1, .., k+1$$

For a normal-gamma model, a theorem from [55, 161] gives that the marginals over σ^2 are given by a multivariate Student t-distribution with n_i dimensions and s_i degrees of freedom:

$$\mathbf{y}_{R_i} \mid k, \tau^{(k)} \sim St_{d_i}(\mathbf{x}_i'\theta_i, \frac{s_i}{d_i}(p_i\mathbf{x}_i'\mathbf{x}_i + I)) \tag{4.17}$$

$$i = 1, ..., k+1$$

Writing the marginal likelihood as a product of the probabilities within each region we have:

$$p(\mathbf{y} \mid k, \tau^{(k)}) = \Pi_{i=1}^{k+1} p(\mathbf{y}_{R_i} \mid k, \tau^{(k)}) \tag{4.18}$$

where

$$p(\mathbf{y}_{R_i} \mid k, \tau^{(k)}) = \frac{\Gamma(\frac{d_i+n_i}{2})(1 + \frac{1}{d_i}(\mathbf{y}_{R_i} - \mathbf{x}_i\theta_i)'\Sigma_i^{-1}(\mathbf{y}_{R_i} - \mathbf{x}_i\theta_i))^{-\frac{d_i+n_i}{2}}}{\Gamma(\frac{d_i}{2})(d_i\pi)^{\frac{n_i}{2}} \mid \Sigma_i \mid^{\frac{1}{2}}} \tag{4.19}$$

and

$$\Sigma_i = (I + p_i\mathbf{x}_i'\mathbf{x}_i)\frac{s_i}{d_i} \tag{4.20}$$

Furthermore, the multivariate Student-t probability density function simplifies using the following a result from [161], greatly increasing computational efficiency. Given observations $y_1, ..., y_n$ i.i.d as $N(\mu_i, \phi_i)$ with the mean and variance unknown and using the conjugate prior, then

$$p(\mathbf{y}_{R_i} \mid \tau^{(k)}) = \frac{(2\pi)^{-\frac{n_i}{2}}(n_i + p_i^{-1})^{-\frac{1}{2}}\Gamma\left(\frac{d_i+n_i}{2}\right)\left(\frac{s_i}{2}\right)^{d_i/2}}{\left(\frac{s_i+z_i}{2}\right)^{\frac{(d_i+n_i)}{2}}\Gamma\left(\frac{d_i}{2}\right)} \tag{4.21}$$

where

$$z_i = \sum_{j=1}^{n_i}(y_j - \bar{y}_{R_i})^2 + \frac{n_i(\bar{y}_{R_i} - \theta_i)^2}{p_i n_i + 1} \tag{4.22}$$

and $\bar{y}_{R_i}$ is the mean of $\mathbf{y}$ within region i. Although it is possible to calculate the posterior for k analytically, it is computationally very intensive. Therefore, we prefer to sample from $p(k \mid \mathbf{y})$ with a Metropolis Hastings algorithm, using that:

$$\begin{aligned} p(k \mid \mathbf{y}) &\propto p(\mathbf{y} \mid \tau^{(k)}, k)p(\tau^{(k)}|k)p(k) \\ &\propto \prod_{i=1}^{k+1} p(\mathbf{y}_{R_i} \mid \tau^{(k)}, k)p(\tau^{(k)}|k)p(k) \end{aligned} \tag{4.23}$$

Posterior Distribution given k and $\tau^{\mathbf{k}}$

Within a particular region, i, we have:

$$\sigma_i^2 \mid \mathbf{y}_{R_i} \sim IG(\frac{d_i^*}{2}, \frac{s_i^*}{2}) \tag{4.24}$$

$$\text{where } d_i^* = d_i + n_i$$

$$s_i^* = s + \frac{1}{2}(\mathbf{y}_{R_i} - \mathbf{x}_{R_i}\theta_i)I^{-1}(\mathbf{y}_{R_i} - \mathbf{x}_{R_i}\theta_i)$$

Since $\mathbf{x}_{R_i}$ is a vector of 1's, s_i^* simplifies:

$$s_i^* = s_i + \frac{1}{2}\sum_{j=1}^{n_i}(y_j - \theta_i)^2 \tag{4.25}$$

The posterior distribution for μ_i is given by the following:

$$\mu_i \mid \sigma_i^2, \mathbf{y}_{R_i} \sim N(m_i^*, V_i^*) \tag{4.26}$$

$$V_i^* = \frac{\sigma_i^2 p_i}{1 + n_i p_i}$$

$$m_i^* = \frac{V_i^*}{\sigma_i^2}(\frac{\theta_i}{p_i} + \mathbf{x}'_{R_i}\mathbf{y}_{R_i})$$

Similarly, m^* simplifies:

$$m_i^* = \frac{\theta_i + p_i \sum_{j=1}^{n_i} y_j}{1 + n_i p_i} \tag{4.27}$$

4.2.1 Metropolis-Hastings Algorithm for Multiple Change Points

The following algorithm is based on a modified version of the sampling procedure outlined in [162]. This is a symmetric procedure so avoids the added complexity of calculating the proposal correction.

At iteration t, given k^{t-1} and τ^{t-1}...

1. Draw a candidate point γ from the possible change point locations, $\{2, .., n-1\}$.

2. If not a change point, γ becomes a change point and k increments by 1 to k^*, if already a change point then remove it and $k^* = k - 1$. Define τ^* as the new configuration of change points including the alteration to γ. Calculate the acceptance ratio using Equation 4.28 and update the current value of k and τ accordingly.

3. Draw a new random permutation of k change points, defining $k^* = k$ and τ^* is the new change point configuration. Calculate acceptance ratio using Equation 4.28 with the new k and τ of the previous step and update to k^t and τ^t accordingly.

The acceptance ratio for an update from k and τ to a proposed k^* and τ^* is defined as:

$$r = min\{1, \frac{p(y|\tau^*, k^*)p(k^*)p(\tau^*|k^*)}{p(y|\tau, k)p(k)p(\tau|k)}\} \tag{4.28}$$

where

$$(\tau^t, k^t) = \begin{cases} (\tau^*, k^*) & \text{with probability } r \\ (\tau^{t-1}, k^{t-1}) & \text{otherwise.} \end{cases} \tag{4.29}$$

Bayes Factor

The Bayes Factor, the ratio of the posterior and prior odds of a change point for each data point, say z, is a useful way of indicating possible change points. We test the hypothesis of each point being a change point against the null, which is not being a change point. Thresholding (using typically a value of 10 or larger) the Bayes Factor (BF) of all data points will leave us with the most likely positions of change points. The marginal prior probability is the same for all possible change point locations and calculated using:

$$p(z \in \tau^{(k)}) = \sum_{i=0}^{n} (\frac{n!}{i!(n-i)!})^{-1} (\frac{(n-1)!}{(i-1)!(n-1-(i-1))!}) \frac{p(i,q)}{\int_0^n p(x,q)dx} \tag{4.30}$$
$$= \frac{q}{n \int_0^n p(x,q)dx} \sum_{i=0}^{n} i(1-q)^i$$

where $p(k, q)$ is the geometric prior density function on k, with parameter q. Using the following relations:

$$\text{prior odds} = \frac{p(z \in \tau^{(k)})}{1 - p(z \in \tau^{(k)})} \tag{4.31}$$
$$\text{posterior odds} = \frac{p(z \in \tau^{(k)} \mid \mathbf{y})}{1 - p(z \in \tau^{(k)} \mid \mathbf{y})} \tag{4.32}$$

where the marginal posterior probability of a point being a change point is computed from the Metropolis Hastings sampling output, the Bayes Factor can be calculated for each data point.

4.2.2 Testing and Convergence

A test dataset was constructed by generating data from Normal distributions with different parameters for five different segments. Running 10000 iterations (5000 burn-in) of the BPM MCMC algorithm and sampling from the posterior distributions of μ and σ every 10th iteration correctly identifies the four change points as well as the mean of each segment as illustrated by Figure 4.4.

Convergence of the algorithm was investigated by varying the MCMC and prior parameters as well as the information contained in the data and was found to be robust to all these factors. Figures 4.5, 4.6 and 4.7 show convergence of the change points for three different test data sets: a noisy signal, a signal with fewer data-points and a noisy signal with few data-points (generated similarly to as described above). The third scenario required slightly more iterations

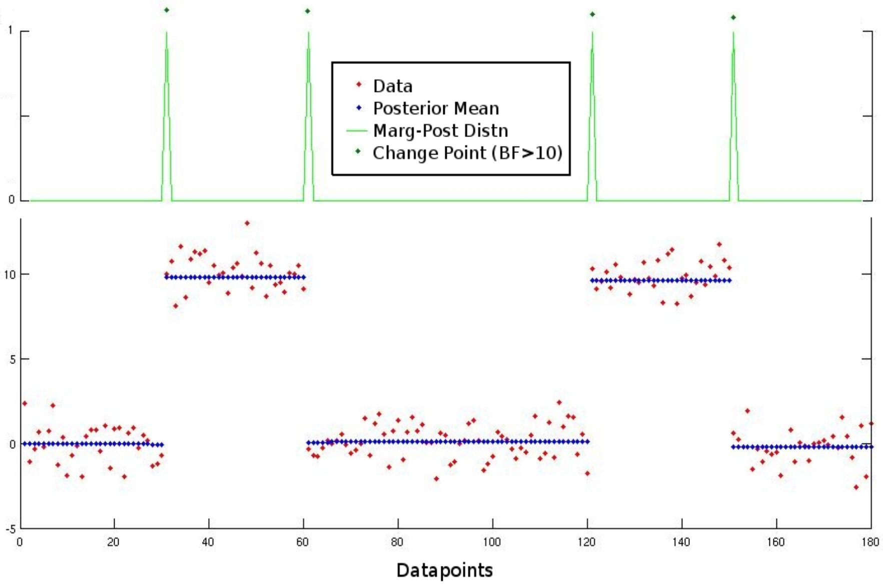


Figure 4.4: Plot of test data (red), posterior mean of distribution within each partition (blue), marginal posterior probability of each data point being a change point (green line) and those points with a Bayes Factor >10are represented by dots at the top of the diagram.

for stability but all three cases converged relatively quickly to the correct answer. Sensitivity of convergence to prior parameters was minimal on the test data and is explored further in the next section using a 'real' signal.

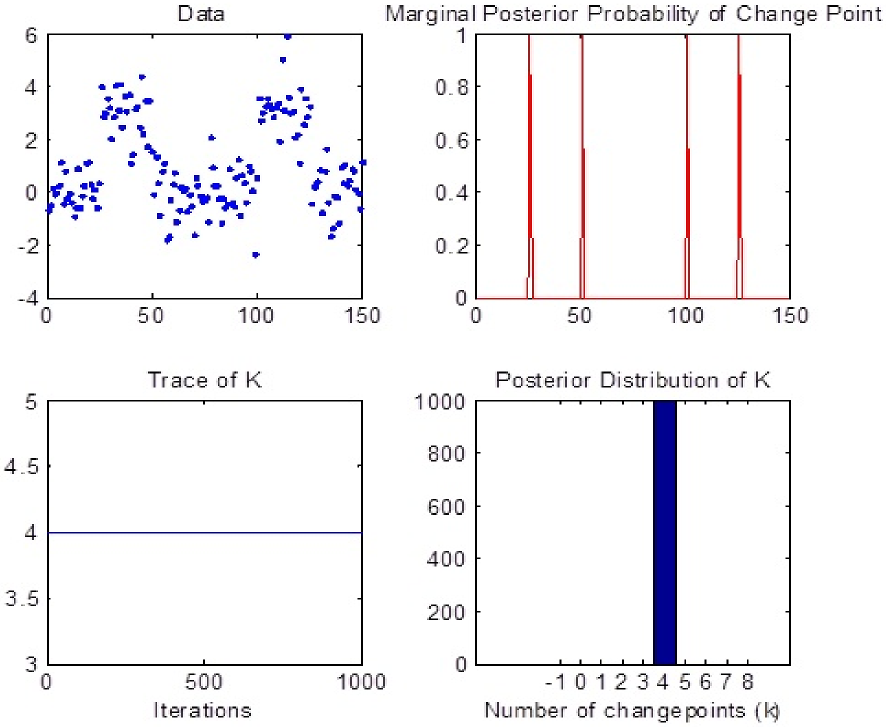


Figure 4.5: Figure shows the signal data modelled (blue scatter plot), the marginal posterior probability of each data-point being a change point (red line plot) correctly indicating the change points, trace plot of K over the run (blue line plot) and the posterior distribution of K (blue histogram). The run was 6,000 iterations long including 5,000 burn-in. Prior mean on θ was 20 and initial value of K was set to 20. Prior parameters p and q were set to 0.5 and 20 respectively and prior parameters d and s were set to 50 and 20 respectively. Despite the data being noisy, the model converges quickly to the correct answer.

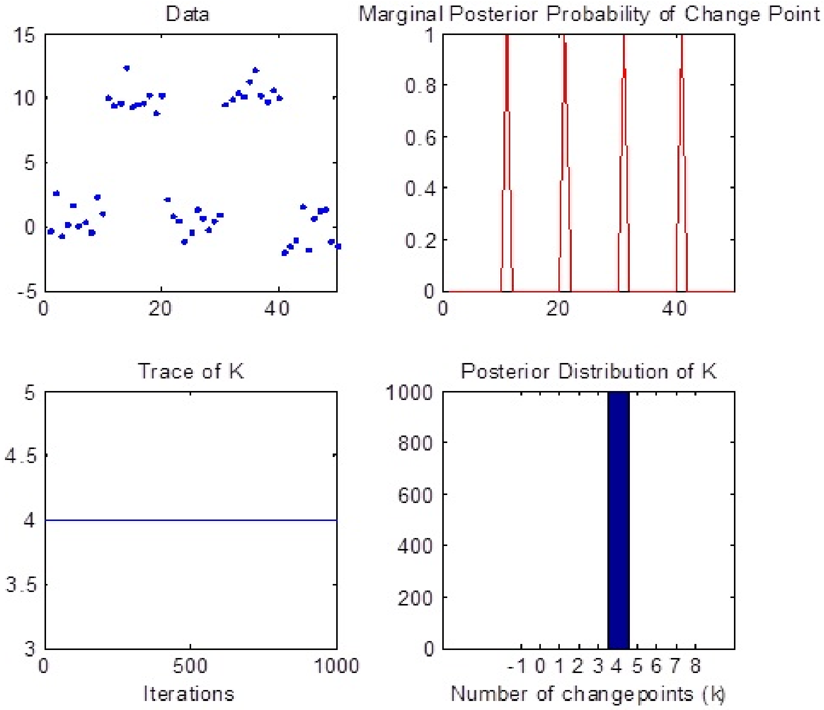


Figure 4.6: Figure shows the signal data modelled (blue scatter plot), the marginal posterior probability of each data-point being a change point (red line plot) correctly indicating the change points, trace plot of K over the run (blue line plot) and the posterior distribution of K (blue histogram). The run was 6,000 iterations long including 5,000 burn-in. Prior mean on θ was 20 and initial value of K was set to 20. Prior parameters p and q were set to 0.5 and 20 respectively and prior parameters d and s were set to 50 and 20 respectively. Despite there being relatively few data-points, the model again converges quickly to the correct answer.

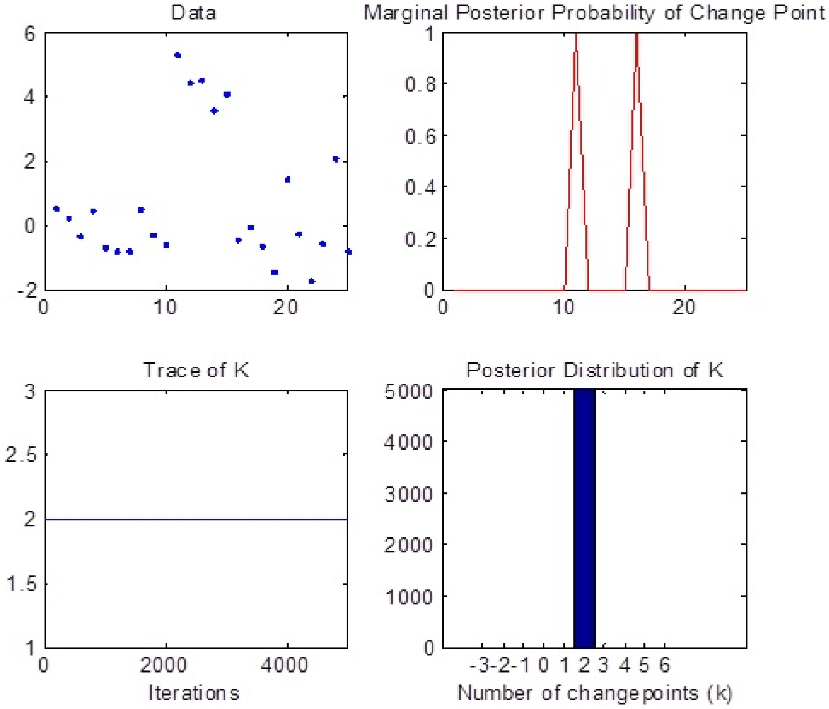


Figure 4.7: Figure shows the signal data modelled (blue scatter plot), the marginal posterior probability of each data-point being a change point (red line plot) correctly indicating the change points, trace plot of K over the run (blue line plot) and the posterior distribution of K (blue histogram). The run was 10,000 iterations long including 5,000 burn-in. Prior mean on θ was 20 and initial value of K was set to 20. Prior parameters p and q were set to 0.5 and 20 respectively and prior parameters d and s were set to 50 and 20 respectively. Given there was considerable noise and few data-points, the model took slightly longer to converge but still did so relatively quickly and to the correct answer.

4.2.3 Sensitivity Analysis

Varying the parameters for the BPM gives us some control over the sensitivity of the peak detection. p is related to the variance of the prior mean for each data segment. Figure 4.8(a) shows that as we allow greater variance of the mean level, θ, we fit a larger number of change points. q is the prior parameter on k the number of change points. The prior mean for k is given by $\frac{1-q}{q}$. As expected, see Figure 4.8(b), setting a low prior mean penalises large k, so fewer change points are detected.

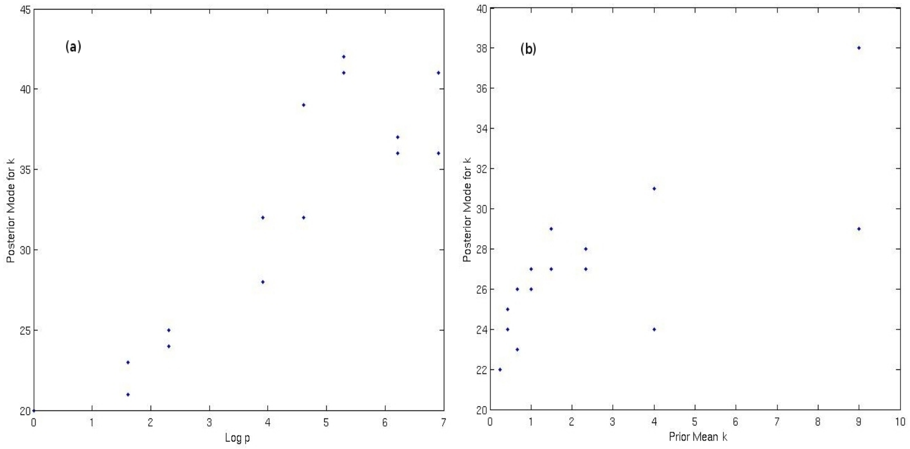


Figure 4.8: Sensitivity Analysis for prior parameter (a) p and (b) q on k indicating that increasing variance for prior on k allows greater number of change points to be fitted.

s and d are the scale and shape parameters, respectively, of the Inverse-Gamma (IG) prior on the variance, σ. The effects of varying these parameters is more complex to gauge since the mean and variance of the IG distribution are $\frac{1}{s(d-1)}$ and $\frac{1}{s^2(d-1)^2(d-2)}$. Figure 4.9 suggest that both parameters are approximately inversely proportional to the number of change points detected. Increasing the value for s or d significantly means that the mean and variance both increase, allowing large variability within each data segment and thus fitting fewer change points.

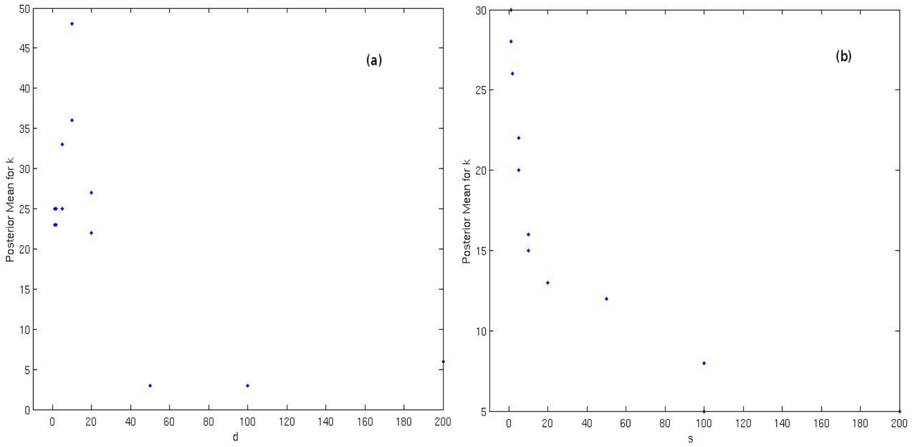


Figure 4.9: Sensitivity Analysis for prior parameters (a) d and (b) s indicating that increases of d or s correspond to higher mean and variance allowing greater variability of intensity within each partition and so fitting fewer change points.

4.2.4 Results

The BPM was run on a 'test' chromatogram at 264.95 m/z (0.1m/z bin), using parameters penalising large values of k. Figure 4.10 shows that the algorithm is capable of estimating the mean and position of peaks fairly accurately. However, the shape of the signal is not well fitted and is an area that needs further investigation. Using a Gaussian or Exponentially Modified Gaussian shaped design vector, $\mathbf{x}$, may model the peaks more precisely. Fixing $k = 2$ and requiring the centre section to be greatly different from zero provides a robust method of peak detection, but is very computationally intense so an alternative strategy is pursued in the next section as part of the H-MS algorithm.

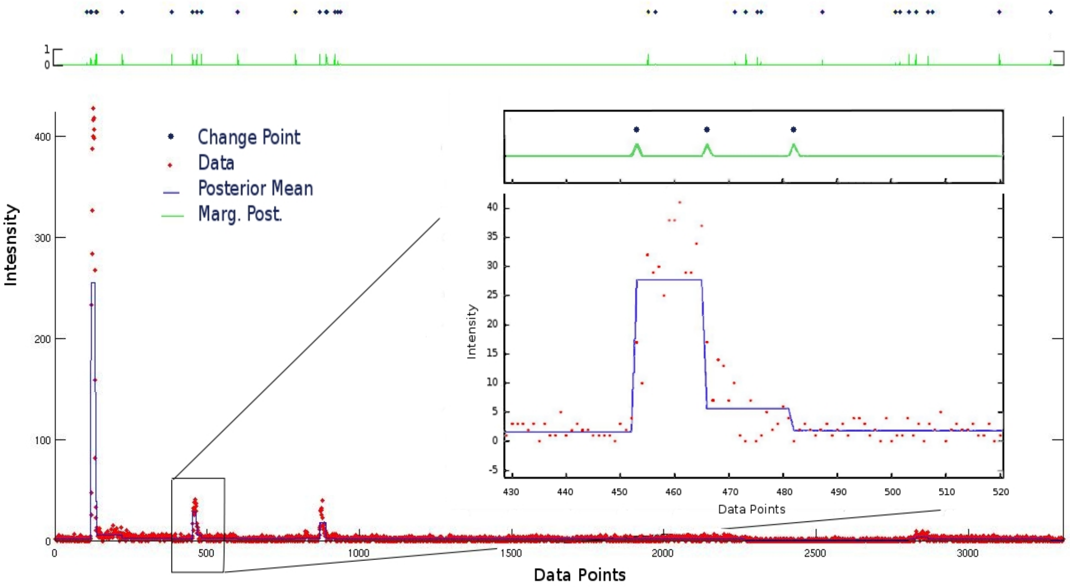


Figure 4.10: Shows the real data (red) overlaid with the posterior mean (blue) and the marginal probability of a change point (green). Blue dots indicate points with a Bayes Factor >10. The inset shows a close-up of a relatively low intensity peak.

4.3 H-MS: Wavelet Smoothing, Adaptive Binning and Peak Detection

Our approach entails a combination of different methodology. H-MS consists of two main algorithms: the first bins the data in the m/z dimension and the second performs peak detection in the Rt dimension within each bin sequentially. This marrying of methods helps ensure that information is not lost by splitting peaks or including unnecessary noise as is often the case when using more primitive binning strategies. The peak detection fits chromatographic peaks with a Gaussian or Exponentially Modified Gaussian (EMG) shape. A number of tests are then used to determine whether a candidate peak is a ‘true’ peak. The methods are outlined in more detail in this section, and their performance is assessed and compared to existing packages apLC-MS and XCMS.

4.3.1 Binning in m/z Dimension

1. Create Integrated Mass Spectrum The whole dataset for a single sample (consisting of m/z, Rt and intensity triples) is ordered according to m/z value, and where duplicate m/z values occur the intensity is summed. This gives us a 'Total Mass Spectrum' (TMS) that will retain features of the m/z dimension of the data.

2. Smooth Spectrum

The spectrum is then transformed using the discrete wavelet transform with Daubechies 3 wavelets (commonly used to model m/z peaks [163]) as default, although other wavelets are available in the MATLAB Wavelet Toolbox. The wavelet coefficients are then thresholded using the universal threshold (based on the median average deviance) method with the option to use the level dependent version [163]. H-MS has the option to perform either 'hard' or 'soft' thresholding, with the default being soft. By performing the inverse wavelet transform, we can see the smoothed spectrum as in Figure 4.11. The wavelet transforms are calculated using inbuilt MATLAB functions within the Wavelet Toolbox.

3. Bin Placement

Local maxima are located using a simple search. The algorithm then descends down either side of the maximum to find the peak edges. The peak edges are defined as the position at which the TMS signal reaches zero, or at which the signal begins to increase again. These peak 'edges' are then used for bin placement.

4. Filter Bins

Bins are then filtered according to a user specified 'maximum peak width'. If adjacent bins both contain possible peaks, the position of each maximum is found. If the distance between the peaks is less than the maximum peak-width, the bins are combined. Adjacent empty bins are amalgamated into a single bin.

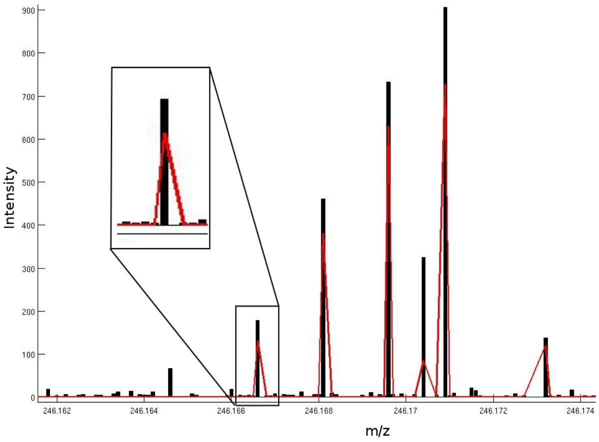


Figure 4.11: An example Total Mass Spectrum (with zoomed inset) showing raw data in black and the wavelet-smoothed data in red.

4.3.2 Peak Detection in Chromatographic Dimension

The Guassian (three free parameters) and Exponentially Modified Gaussian (four free parameters) are widely considered good models for chromatographic peak shape [164]. The Gaussian model is described by the following equation:

$$f(x) = Ae^{-\frac{(x-b)^2}{2c^2}} \tag{4.33}$$

where b is the center of the peak, c^2 is the variance and A is the amplitude. The Exponentially Modified Gaussian (EMG) model is given by the equation:

$$EMG(x) = \frac{A}{\tau\sigma(2\pi)^{1/2}} \int_0^x \exp(-\frac{(x-b-x')^2}{2c2}) \exp(-\frac{x'}{\tau}) dx' \tag{4.34}$$

where b is the centroid maximum of the Gaussian component, c^2 is the variance of the Gaussian component, τ is the time constant of the exponential function and A is the amplitude. The H-MS peak detection routine employs both these models in order to accurately capture differing peak shapes.

For each bin, a chromatogram is constructed containing the data in the bin across the range of retention times. To detect peaks within each chromatogram, the maximum within the slice is identified and the candidate peak area is defined within a given window of the maximum. A number of tests are carried out within the following framework, designed to identify well-shaped peaks:

1. Fit Gaussian Model

The Gaussian model is fitted to the candidate peak area using least squares. The peak beginning and end are calculated as the points at which the intensity of the fit falls below either a given proportion of the maximum intensity (default is 5 %) or a simple threshold.

2. Peak Width Check

Candidate peaks with widths (in Rt) below a specified 'minimum peak width' and above a 'maximum peak width' are discarded.

3. Data-points Check

Peaks with fewer than a specified number of non-zero data-points are removed.

4. Likelihood Ratio Test

The Gaussian peak model is compared to the hypothesis of zero signal (described in further detail in the next section).

5. Smoothing Spline Shape Test

The Gaussian peak model is compared to the fit provided by a smoothing spline in order to reject poorly shaped peaks (described in further detail in the next section). Candidate peaks that pass this test are considered 'true' peaks.

6. Fit Exponentially Modified Gaussian Model

If the candidate peak is rejected by Step 5, we consider the possibility that the peak may be better characterized by the EMG model so Steps 2-5 are repeated using the new fit. Employing the EMG guards against rejecting true peaks with large tails that are fitted poorly by the standard Gaussian.

7. Local Minima Check

If, after trying the EMG fit, the candidate peak is still rejected at Step 5, we check for the presence of more than one peak within our candidate peak area by searching for a local minima. The smoothing spline is re-fit but this time with less smoothing, in order to avoid local minima arising from noisy data but retain the trough between two separate peaks. If a local minima is found, the candidate peak area is split at the local minima resulting in a new candidate peak. Then, the complete peak detection procedure (from Step 1) is repeated for this new candidate.

If a peak is detected, then the fitted parameters are added to the peak list and the intensity between peak start and end is set to zero so as not to interfere with detection of further peaks within the same slice. If no peak is detected, the candidate section of signal is discarded as noise and set to zero. The algorithm then looks at the next maxima in the chromatogram and checks for further peaks. The peak m/z value is calculated by taking an intensity-weighted average of the raw data-points within the detected peak edges. The peak Rt is given by the value at which the peak attains its maximum intensity. The peak intensity is calculated by summing the intensity of the raw data points within the peak edges. Once there are no further maxima in the chromatogram above the given threshold, the algorithm moves on to the chromatogram of the next m/z bin.

Likelihood Ratio Test for Presence of Peak

The likelihood ratio for the least squares fit (alternative hypothesis) versus no peak being present (null hypothesis) is calculated and accepted/rejected with 5% significance:

$$L = \frac{p(y, \mu_1, \sigma_1^2)}{p(y, \mu_0, \sigma_0^2)} \tag{4.35}$$

$$-2ln(L) \sim \chi^2_{df} \tag{4.36}$$

where y is the data, $p(y, \mu_i, \sigma_i^2)$ is the Normal likelihood of a particular model, μ_1 is the least squares estimate of either the Gaussian or EMG model, σ_1^2 is the variance of the residuals of this estimate, μ_0 is a vector of zeros and σ_0^2 is the variance of the residuals from the null hypothesis.

$-2ln(L)$ then has a χ^2 distribution with 3 (when using the Gaussian fit) or 4 (when using the EMG fit) degrees of freedom. The degrees of freedom, df, is calculated as difference between the number of free parameters of the alternative model and that of the null which has one free parameter.

Smoothing Spline Shape Test

Here we use an inbuilt MATLAB smoothing spline function to model the data as the null hypothesis, with our Gaussian or EMG fit as the alternative. The Bayesian Information Criterion (see Section 3.1.2 for more details on this method of model selection) is then used to choose between the two models since it will penalise the large number of parameters of the smoothing spline (calculated by a MATLAB fitting routine). It is hoped that for poorly shaped candidate peaks the spline fit will be preferred, despite the number of parameters, so the peak is rejected.

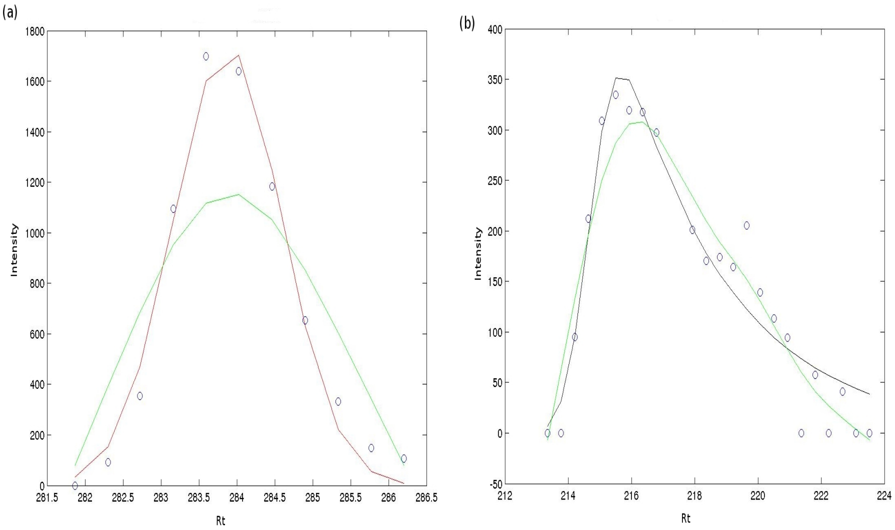


Figure 4.12: Candidate peaks accepted by smoothing spline shape test using (a) Gaussian model and (b) Exponentially Modified Gaussian model. Blue circles indicate raw data points, red line indicates Gaussian fit, black line indicates EMG fit and green line indicates smoothing spline fit.

Figure 4.12 shows (a) an example of a peak detected using the Gaussian model and (b) an example of a peak detected using the EMG model after being rejected using the Gaussian, reinforcing the benefits of employing both models.

4.3.3 Results and Discussion

We use a Synthetic Human Urine Sample, a Standard Dilution Series (SDS) of human urine metabolites and simulated 'noise' datasets to explore the binning and peak detection capabilities of H-MS. Given the variability of LC-MS datasets it was important to use multiple sets for evaluating robustness and also the data-set selection gives scope to evaluate different characteristics of the algorithms.

The Synthetic Human Urine Sample contains 83 different metabolites and the resulting largely unpredictable interactions and noise should provide a good simulation of the complexities of data arising from a real organic biofluid whilst allowing us to have prior knowledge of the sample composition to aid peak searching and identification. Whilst using a real urine sample where the composition is known would be the gold standard this is obviously impossible so analysis of a realistic synthetic sample will allow evaluation of the algorithms in as close to a 'real-life' setting as possible where data is noisy and complex, whilst knowing the approximate 'true answer'.

The Standard Dilution Series is a less complex mixture (14 metabolites) and as such we would expect to see a simpler dataset set where peak detection and quantification would perhaps be easier than for the synthetic urine sample certainly for high concentrations. The dilution series of each metabolite should exhibit a linear trend in terms of concentration levels so examining the peak sizes captured by the algorithms should give us a good indication of how well the peaks are characterised. Deviations from linearity may highlight peaks that are either poorly characterised (i.e. peaks are split or multiple peaks are grouped together) or peaks that do not belong to the metabolite, i.e. false positives, that are much harder to identify in the synthetic urine sample. This analysis should also show how well the algorithms distinguish between noise and analyte peaks at lower concentrations.

We compare H-MS to the adaptive binning and peak detection algorithm apLC-MS [66] and both the centWave and matchedFilter peak detection routines of the commonly-used package XCMS [130, 46]. XCMS was chosen given its prominent use across the field of metabolomics

([46] has 634 citations) and the fact that it offers two different methods for peak detection, one of which utilises wavelet analysis in an alternative way to H-MS. apLCMS was selected for its claim to effectively perform adaptive binning comparative to that of H-MS. As shown in Section 4.1.2 there are a multitude of algorithms that could be compared, but for the purposes of this section those selected give a good balance of a more traditional filtering method in XCMS' matched filter, a robust and highly regarded 'gold standard' in XCMS' centWave and a perhaps more innovative and unconventional method in apLCMS.

As previously mentioned, centWave uses the Continuous Wavelet Transform (CWT) to locate chromatographic peaks on different scales and identify prominent features, whereas matchedFilter uses a Mexican Hat-shaped filter to detect peaks across the Rt dimension within fixed-width m/z slices of the data.

apLC-MS performs adaptive binning by searching for m/z tolerance level using a mixture model and then grouping m/z values based on the tolerance level. The mixture model is based on the assumption that the consecutive differences between values of an ordered list of the detected m/z for a sample consist of differences within the same peak (caused by small measurement variations) and differences between peaks and noise data-points. The differences between peaks or noise data-points are approximated by the spacings between a sample drawn from a Uniform distribution which can be modelled using an exponential distribution, whilst the differences of m/z values within a single peak are modelled using an unspecified function with a maximum value that is very small. The m/z tolerance level is then based on the estimated parameters of the mixture model. The grouped data points are then refined using non-parametric density estimation in both m/z dimension and elution time dimension, with groups split at valley positions if multiple models are identified. A run filter is also applied, requiring a peak to have a minimum length in the retention time dimension, as well as being detected among at least the specified proportion of the time points within the time period. After this procedure, features are fitted with a kernel smoother to determine the peak parameters, using an Expectation Maximisation algorithm with pseudo likelihood for features with multiple peaks.

Similarly to XCMS, apLC-MS also incorporates a Retention Time alignment algorithm but as mentioned previously, exploring these techniques is beyond the scope of this thesis and this pre-processing step will be ignored in order to perform a fairer comparison and evaluation of the algorithms peak detection capabilities.

Data

Two experimental data-sets were used to evaluate the algorithms.

Synthetic Urine Sample

Preparation

Eighty-three of the most abundant endogenous mammalian metabolites, ranging in molecular weight from 30-625 Da, were weighed into a 1-L bottle and then dissolved in 1L of HPLC-grade water (Sigma-Aldrich, St. Louis, MO). Any remaining solids were removed by vacuum filtration. Approximate final metabolite concentrations were targeted to fall between 1mM and 20mM, with sodium azide added at 0.05% v/v as a preservative. The normally high levels of inorganic salts found in urine were not added, in order to eliminate the effect of salt suppression in the various sample introduction interfaces. The stock solution was stored at -80°C.

Instrumentation

Synthetic urine samples (5μL) were injected onto a 2.1mm × 100mm (1.7μm) HSS T3 Acquity column (Waters Corporation, Milford, CT) and eluted using a 18-min gradient of 100% A (water, 0.1% formic acid) to 100% B (acetonitrile, 0.1% formic acid). The flow rate was 500μL/min, the column temperature was 40°C, and the sample temperature was 4°C. Samples were analysed using a UPLC system (UPLC Acquity, Waters Ltd. Elstree, U.K.) coupled online to a Q-TOF Premier mass spectrometer (Waters MS Technologies, Ltd., Manchester, U.K.) in positive- and negative-ion electrospray mode with a scan range of m/z 50-1000 and a scan time of 0.08s. Three technical replicates were run. To obtain data that were as raw as possible, the spectrometer was run in continuum mode and the detector saturation correction was switched off. A feature of the Q-TOF Premier is that it employs a 'DRE lens', which is a mechanism for defocusing the ion beam to minimize detector saturation. This defocusing mechanism was also switched off.

Tables 4.3.3 and 4.3.3 list the metabolites included along with their approximate concentration and indicate whether they were manually identified in the LC-MS (positive and negative runs) by the group who ran the experiments and also if they were detected automatically by the algorithms evaluated.

Standard Dilution Series

Preparation

A total of 14 metabolites commonly observed in human urine were chosen as standard compounds and obtained in stable isotope (^{2}H or ^{13}C) labelled form. All samples were prepared in a single batch at Imperial College London. All reagents including LC-MS grade water, acetronile with pre-added 0.1% formic acid, and standard compounds were obtained from Sigma Aldrich (Gillingham, UK). Columns and maximum recovery vials were donated by Waters Corporation (Milford,US). Stock mixtures of the standards (1mg/mL) were prepared and diluted to a working concentration of approximately 10μg/mL per standard. These stock solutions were subject to four two-fold dilutions resulting in dilutions of 1, 1/2, 1/4, 1/8 and 1/16.

Instrumentation

Synthetic urine samples were analysed using a UPLC system (UPLC Acquity, Waters Ltd. Elstree, UK) coupled online to a Q-TOF Premier mass spectrometer (Waters MS Technologies, Ltd., Manchester, U.K.) in positive electrospray (ESI) mode with a scan range of 50-1000 m/z. Capillary voltage was 2.4Kv, sample cone was 35V, desolvation temperature 350°C, source temperature 120°C, and desolvation gas flow 900 L/hr. The Q-TOF Premier was operated in V optics mode, with a data acquisition rate of 0.1s and a 0.01s inter-scan delay. Leucine enkephalin (m/z 556.2771) was used as the lockmass; a solution of 200pg/μl (50:50 ACN:H2O) was infused into the instrument at 3μl/min via an auxillary sprayer. Data were collected in centroid mode with a scan range of 50-1000 m/z, with lockmass scans collected every 15s and averaged over 3 scans to perform mass correction.

Name	Concentration (mM)	Manually Identified	Automatically Detected
1-methylnicotinamide	1.7		
3-(3-hydroxyphenyl)propionic acid	9.21		
3-hydroxycinnamic acid	1.18	X	X
3-methyladipic acid	23.73		
3-nitro tyrosine	0.13	X	
4-Aminobutanoic acid	10.09		
4-aminohippuric acid	11.07	X	X
4-hydroxyproline	1.14		
4-methyl-2oxopentanoic acid	9.68		
5-hydroxyindole acetic acid	1.57		
acetamide	4.42	Outside scan range	
adenosine triphosphate	0.78	X	
adipic acid	11.63	X	
alanine	1.83	X	
allantoin	0.99		
arabinose	0.74		
ascorbic acid	9.48	X	
betaine	3.28	X	
chenodeoxycholic acid	7.82	X	X
cholesteryl palmitate	0.88		
choline	2.98	X	
cis-aconitic acid	1.67		
citric acid	10.05		
creatine	1.22	X	X
creatinine	1.49	X	
cyclopropanedicarboxylic acid	2.21	X	
D-arginine	0.72		
D-asparagine	0.72		
D-Aspartic acid	10.14		
D-Lactic acid	5.55	X	
D-lysine	2.22		
D-malic acid	13.72	X	
D-ornithine	0.82		
D-Proline	1.01	X	
D-tyrosine	1.17		
ethanolamine phosphate	0.86		
folic acid	7.07		
fumaric acid	18.95	X	
glucuronic acid	7.73		
glutaric acid	10.37		
glycerol	1.16	X	
glycine	4.36	Outside scan range	
glycolic acid	12.75	Outside scan range	
hippuric acid	15.07		

Table 4.2: Chemical composition of synthetic urine sample and approximate concentration of analytes. Metabolites manually identified in LC-MS runs and those automatically detected by at least one algorithm indicated by ‘X’.

Name	Concentration (mM)	Manually Identified	Automatically Detected
homoserine	0.93	X	
hypoxanthine	0.9		
indoxyl sulfate	0.06	X	
isobutyric acid	5.68	X	
isocitric acid	14.6	X	X
L-carnitine	0.9	X	
L-citrulline	0.78	X	X
L-Serine	1.25		
L-Threonine	3.88	X	
L-tryptophan	1.13	X	X
mannitol	1.22		
methyl succinic acid	16.35		
methylamine	6.49	Outside scan range	
methylmalonic acid	1.02	X	
myo-inositol	1.4		
N,N-dimethylbenzamide	1.31		
N-acetyl-L-glutamic acid	9.52		
nicotinic acid	11.7		
N-Methylglycine	1.21	X	
N-methyl-L-histidine	0.14		
oxalacetic acid	11.74		
oxoglutaric acid	9.16		
phosphoenolpyruvic acid	0.07		
pimelic acid	10.11	X	X
propanoic acid	12.79	Outside scan range	
pyruvic acid	5.68	X	
riboflavin	1.05	X	X
salicylic Acid	11.8		
sebacic acid	10.73		
suberic acid	1.95	X	X
succinic acid	5.62	X	
thymine	1.09	X	
trimethylamine N-oxide	1.38	Outside scan range	
uracil	1.12		
uric acid	1.04	X	X
uridine	1.4		
urocanic acid	11.95		
xanthine	1.09	X	X
xylose	1.06		

Table 4.3: continued: Chemical composition of synthetic urine sample and approximate concentration of analytes. Metabolites manually identified in LC-MS runs and those automatically detected by at least one algorithm indicated by ‘X’.

Table 4.3.3 lists the metabolites included, indicating whether they were detected by the algorithms evaluated and if so, in what chemical form.

Name	Detected	Form(s) detected
acetylcarnitine-d3	X	M, M+H, M+Na, 2M+H, 2M+Na
adipic acid-d8	X	M+H
dimethylglycine-d6	X	M+Na
DL-leucine-d3	X	M+H, M+CO2
DL-methionine-13C	X	M+H, M+Na
DL-phenylalanine-13C	X	M+H
dopamine-d4	X	M, M+H
glutaric acid-d4		
heptanedioic-d4	X	M+H
hippuric acid-d2	X	M+H, 2M+H, 2M+Na
L-DOPA-ring-d3	X	M, M+H, M+Na, 2M+H, 2M+Na
nicotinamide-d4	X	M, M+Na
succinic acid-d4	X	M+Na
tryptamine-d4	X	M, M, M+H

Table 4.4: Chemical composition of standards dilution series batch. 'X' indicates detected in any form (including dimers and adducts) in any replicate by any of the algorithms, with the detected forms also listed.

Estimating the False Positive Rate

The null hypothesis that the sample contains no peaks was tested for the H-MS, centWave, matchedFilter and apLC-MS. All the null hypothesis testing was repeated for a variety of parameters for each method.

centWave has four main parameters: peak-width range (default is 20-50s), signal-to-noise threshold (default is 10), ppm - maximum m/z deviation allowed between consecutive scans when identifying initial 'regions of interest' (default is 25ppm), and the pre-filter settings (k,I) where mass traces are only retained if they contain at least 'k' peaks with intensity greater than 'I' (default is $k = 3$, $I = 100$). The peakwidth was varied using ranges of (2,20), (2,50) and (20,50). The signal-to-noise thresholds used were 5,10 and 20. The ppm setting was varied as 10,25 and 50. The prefilter settings used were (3,50), (3,100) and (5,100).

For matchedFilter, the main parameters are the 'full width at half maximum' of the matched filtration Gaussian model peak (default is 30), maximum number of peaks per Extracted Ion

Chromatogram (EIC) (default is 5) and signal-to-noise threshold (default is 10). The full width at half maximum values used were 15,30 and 60. The maximum peaks per EIC were set at 3, 5 and 8. The signal-to-noise threshold was varied using values 5, 10 and 20.

When running apLC-MS binning, the relevant parameters are the run filter parameters, i.e the minimum proportion of presence in the time period for a series of signals grouped by m/z to be considered a peak (default is 0.5) and the minimum length of elution time for a series of grouped signals to be considered a peak (default is 12), and the m/z tolerance level for the grouping of data points expressed as a fraction of the m/z value (default is 1e-5). The minimum proportion of presence was set at 0.3, 0.5 and 0.8. The minimum elution times used were 2, 5 and 12. The m/z tolerance values were 5e-6, 1e-5 and 5e-5.

When running H-MS, three different peakwidth ranges were used: 2-20, 2-50 and 20-50 seconds. Maximum peakwidths in the m/z dimension were given by 0.001, 0.0025 and 0.005 Daltons. The intensity was hard thresholded using thresholds of 50 and 100 counts.

Noisy data was simulated using a 2D Poisson point process (10 replicates with 5 different rates: 0.05, 0.01, 0.005, 0.001 and one estimated from noisy regions of the data equal to 0.0025) since this is the distribution commonly assumed for LC-MS noise [165]. XCMS (both matchedFilter and centWave) and H-MS correctly identified no peaks present. However, apLC-MS detected some peaks for small values of the parameter defining the minimum elution time for a group of data-points to be considered a peak, shown in Table 4.5.

Poisson Process Rate	min,run	min,pres	tol	Mean no. of Peaks	Standard Deviation
Data Estimate	2	0.3	1E-5	18.3	59.8
0.05	2	0.3	1E-5	0.4	0.8
0.01	2	0.3	1E-5	115.4	62.7
0.005	2	0.3	1E-5	46.6	72
0.001	2	0.3	1E-5	5.8	21.5

Table 4.5: Mean and variance of number of peaks detected by apLC-MS over 10 replicates of noise datasets simulated using five different Poisson process rates, for a range of values for parameters min.run (minimum length of elution time for a series of grouped signals to be considered a peak), min.pres (minimum proportion of presence in the time period for a series of signals grouped by m/z to be considered a peak) and tol (m/z tolerance level for the grouping of data points expressed as a fraction of the m/z value).

In addition, 10 'shuffled' data sets were tested on all four methods. The shuffled data consisted of the synthetic urine sample, with both intensity and m/z values randomly permuted for fixed retention times in order to disrupt peak shape but retain data-points with a range of intensities. Again, centWave and H-MS correctly deduced there were no true peaks present, whilst matchedFilter and apLC-MS detected a number of peaks depending on the parameters used, shown in Tables 4.6 and 4.7.

fwhm	max	snthresh	Mean No. of Peaks	Standard Deviation
30	5	10	6245.7	65.8
15	5	10	10436.7	82
60	5	10	3862.1	75.8
30	3	10	6222.8	66.9
30	8	10	6246.7	66.8
30	5	5	11759.5	86.84
30	5	20	2992.4	24.6

Table 4.6: Mean and variance of number of peaks detected by matchedFilter over 10 'shuffled' datasets for a range of values for parameters fwhm (full width at half maximum of Rt peak model), max (maximum number of peaks per Extracted Ion Chromatogram) and snthresh (signal-to-noise threshold).

min.run	min.pres	tol	Mean no. of Peaks	Variance
12	0.5	1E-5	243.2	16.2
5	0.5	1E-5	3144.7	93.4
20	0.5	1E-5	87	9.9
12	0.3	1E-5	5436.2	95.8
12	0.8	1E-5	21.6	4.9
12	0.5	5E-6	40.8	4.6
12	0.5	5E-5	24814.9	177.8
2	0.3	1E-5	100369.8	1243763582

Table 4.7: Mean and variance of no. of peaks detected by apLC-MS over 10 'shuffled' datasets for a range of parameters: min.run (min. length of Rt for a series of grouped signals to be considered a peak), min.pres (min. proportion of presence in the time period for a series of signals grouped by m/z to be considered a peak) and tol (m/z tolerance for the grouping of data points expressed as a fraction of the m/z value).

Comparison using Synthetic Urine Sample

In this section, we compare the results of the H-MS binning strategy for a synthetic urine sample, with those of apLC-MS since it also uses an adaptive method. apLC-MS bin positions were located by using the proc.cdf() function which performs the m/z threshold grouping and applies the run filter. Bin positions were then determined by finding the m/z values at the which the feature group begins/ends.

Figures 4.13 and 4.14 show two examples of H-MS outperforming apLC-MS. Here, apLC-MS has split the Hydroxycinnamic Acid peak down the middle and the Creatine peak into several, whilst H-MS preserves the entirety of the peak profiles. For the rest of the identified peaks while H-MS binning offers precision for the majority and does not split them in the way apLC-MS is shown to, apLC-MS places very large bins that are likely to contain much noise. Further figures have not been included since the bins are too wide for illustration to be informative. Whilst XCMS' matched filter does not split peaks in the same way as apLC-MS, it too provides peak parameters that are too broad to be visually compared with those of centWave and H-MS.

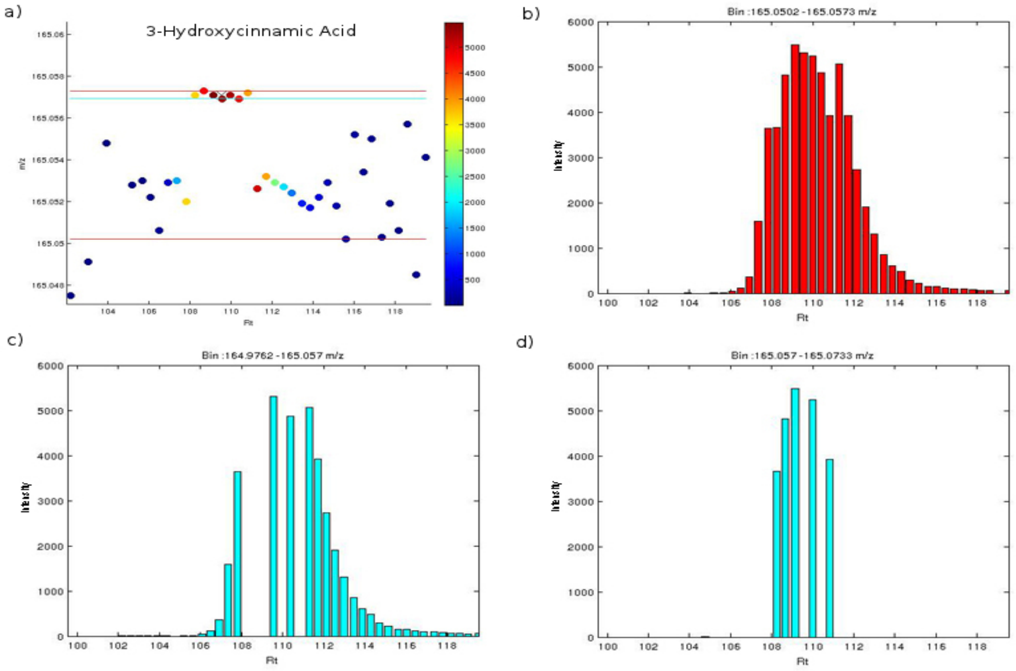


Figure 4.13: A 3-Hydroxycinnamic Acid peak intensity plot (a) of the raw data, with the colour scale indicating intensity and bin edge locations shown for H-MS (red lines) and apLC-MS (cyan lines). Also plotted are chromatograms of the H-MS bin (b) and the two apLC-MS bins (c, d).

In order to evaluate peak detection capabilities, we looked at a number of compounds present in the synthetic urine sample and attempted to cross match them with the peaks detected by each algorithm. Peak identification was performed by allowing a 10ppm × 20 second window about the expected location of the metabolites. centWave was able to identify 17 different

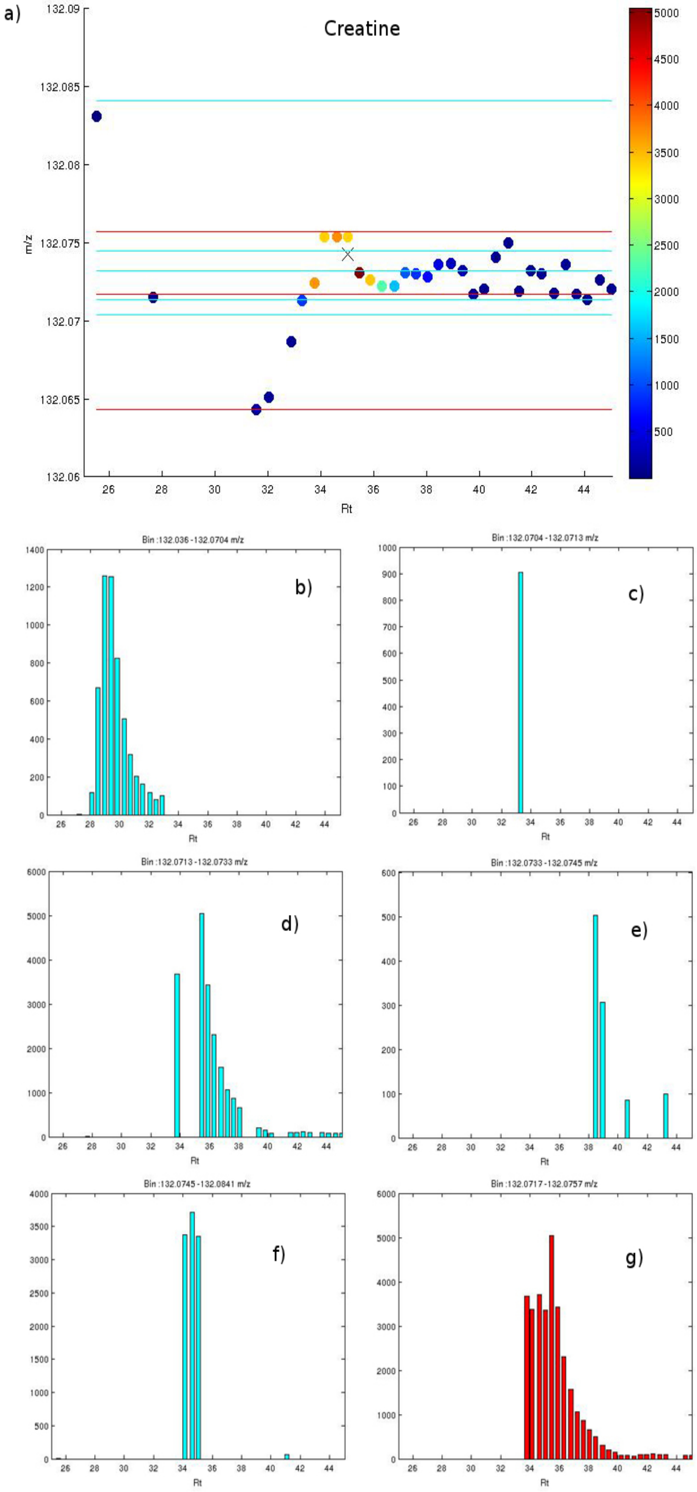


Figure 4.14: A Creatine peak intensity plot (a) of the raw data, with the colour scale indicating intensity and bin edge locations shown for H-MS (red lines) and apLC-MS (cyan lines). Also plotted are chromatograms of the H-MS bin (g) and the five apLC-MS bins (b-f).

features relating to 12 of the metabolites. After varying parameters to get the best results for each algorithm: H-MS and apLC-MS also correctly identified all of these features whilst matchedFilter identified 14.

Since both apLC-MS and matchedFilter give relatively imprecise parameters for detected ‘peaks’ in comparison to H-MS and centWave, we focus on the latter two as the performance is more comparable.

Figure 4.15, shows the raw data intensity plots for six metabolites: 4-Aminohippuric acid, 3-Hydroxycinnamic acid, Pimelic acid, Xanthine, Riboflavin and Isocitric acid with the output peak parameters for both H-MS and XCMS’ centWave. The results from both algorithms are quite similar but H-MS often provides a more accurate description of the peaks. These plots provide some useful illustration of typical peaks, but a more robust evaluation of peak characterisation is described later in this chapter.

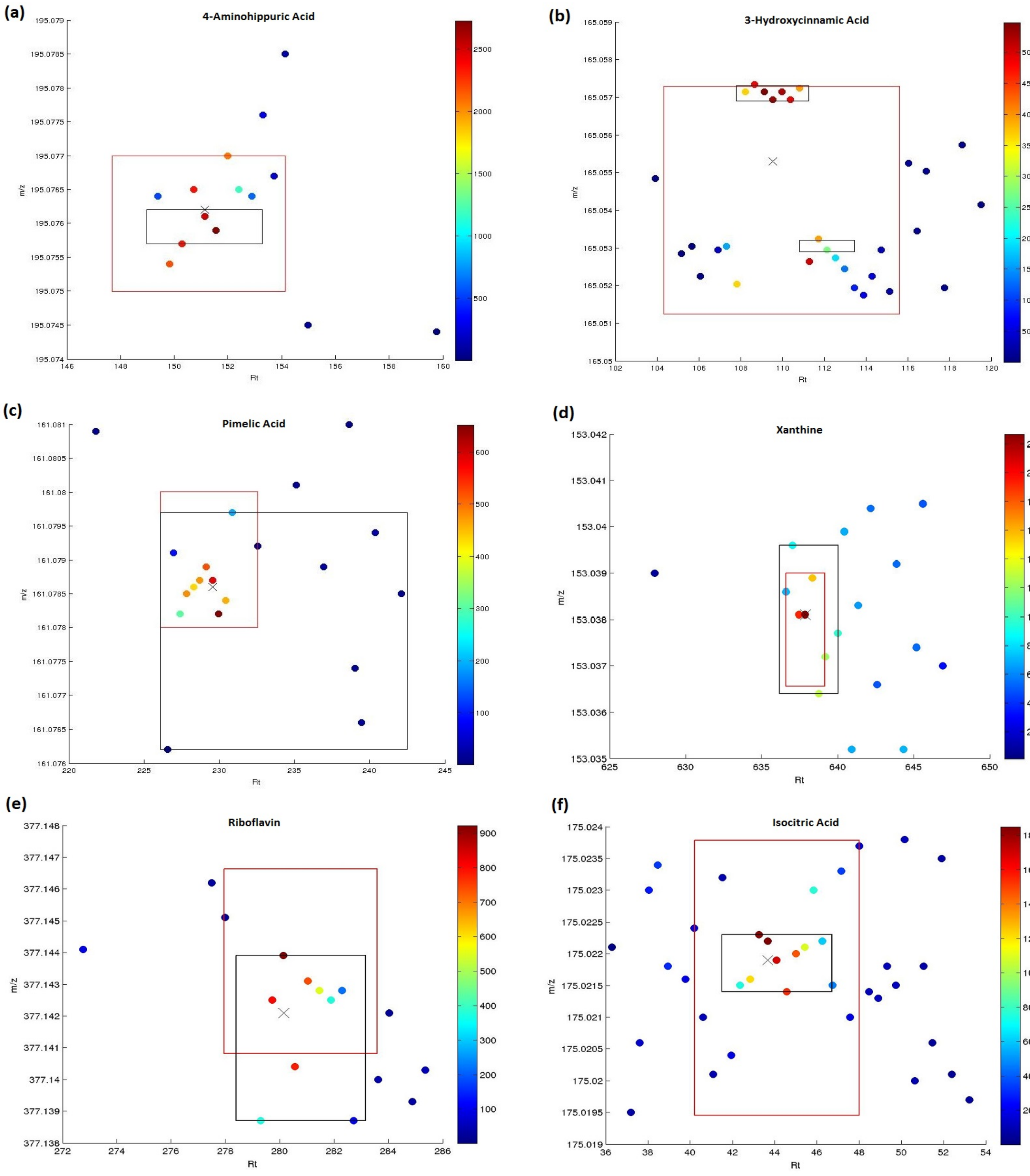


Figure 4.15: Raw intensity plots, with colour scale indicating intensity, of 6 detected metabolite peaks. Peak parameters shown for H-MS (red) and centWave (black) as boxes. (a) 4-Aminohippuric acid gives an example of centWave excluding high intensity data-points but H-MS captures the whole peak. For 3-Hydroxycinnamic acid in (b) we see a substantial drift in m/z value within a peak. H-MS includes all of the peak, whilst centWave detects two separate peaks. (c) Pimelic acid demonstrates where H-MS has defined a tighter area than centWave, including less noise. (d) Xanthine and (e) Riboflavin shows where H-MS excludes some relatively high intensity data-points that are captured by centWave. In (f) H-MS includes some lower intensity datapoints surrounding the Isocitric acid peak which perhaps should not be included as per the centWave parameters.

Comparison using Standard Dilution Series

centWave and H-MS were run on each sample separately (no across-sample peak matching or Retention Time alignment was performed using centWave for a fairer comparison). The resulting peak lists were cross matched with the standard compounds and common adducts and dimers within a 10ppm × 20 second window of the expected peak locations, as determined by a pilot study. Table 4.8 details the average number of peaks detected by each algorithm across the three replicates at each dilution, along with the average number of peaks matched at expected peak locations. The fact that centWave matches more expected peaks could be indicative of greater sensitivity than H-MS, whilst the higher ratio of peaks matched to peaks detected at lower dilutions could suggest H-MS boasts better specificity. However, there are likely to be many unknown peaks expressed in the data (perhaps close to the expected peak locations) so we cannot be certain that peaks have not been mismatched.

Dilution Level	centWave Peak Matches	centWave Total Peaks	H-MS Peak Matches	H-MS Total Peaks
1	18.7	376.0	16.3	311.0
0.5	20.7	329.0	13.7	264.0
0.25	17.7	284.0	15.0	238.7
0.125	11.3	254.3	10.7	195.3
0.0625	10.0	233.7	9.0	169.7

Table 4.8: Average total number of detected and average number of matched peaks for centWave and H-MS algorithms for each dilution level.

Looking at the metabolites (or adducts/dimers thereof) that are detected across all five dilutions for each set of replicates, we can plot the intensity curves for each. Given the dilution series, we expect to find a linear relationship between dilution and intensity, ignoring possible saturation or limit of detection effects. H-MS detects 8 additional peaks consistently across the dilution series, whereas centWave only detects 4. Figure 4.16 shows those peaks detected for both H-MS and centWave, whilst Figures 4.17 (a) and (b) show the peaks detected only by H-MS and only by centWave respectively.

Most of the intensity curves can be well approximated by a linear function as expected, although there are a few exceptions (see Tables 4.9-4.11). Those that did not exhibit a clear linear

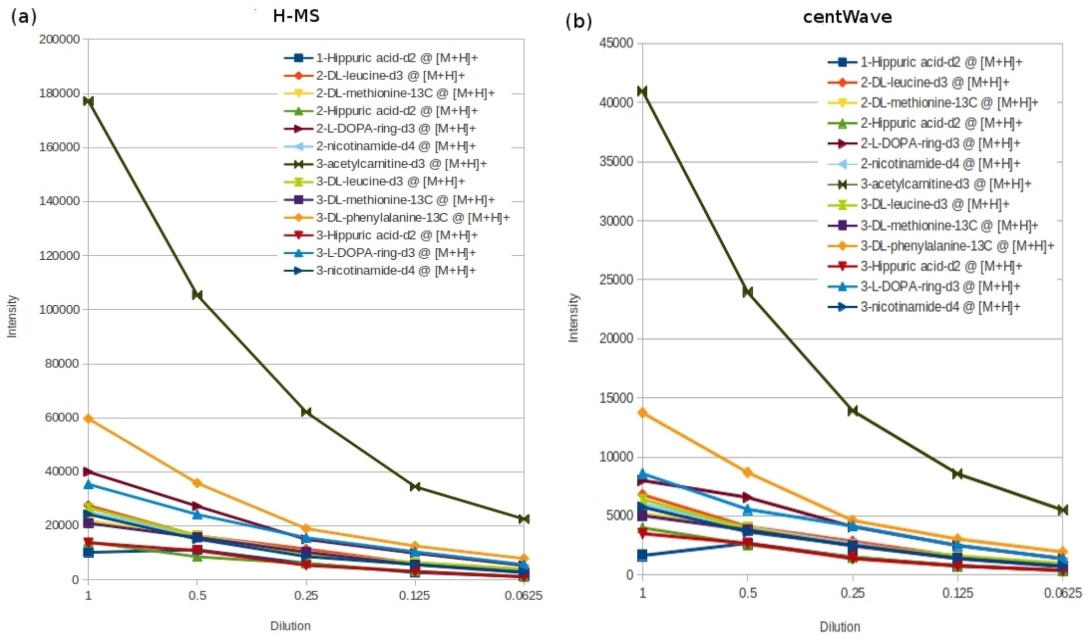


Figure 4.16: Intensity curves for all peaks detected by both a) H-MS and b) centWave across the entire dilution series. The legend gives the molecule names, with the number in front indicating the replicate sample (1, 2 or 3). Strong linear trends give confidence that metabolites have been accurately identified and matched.

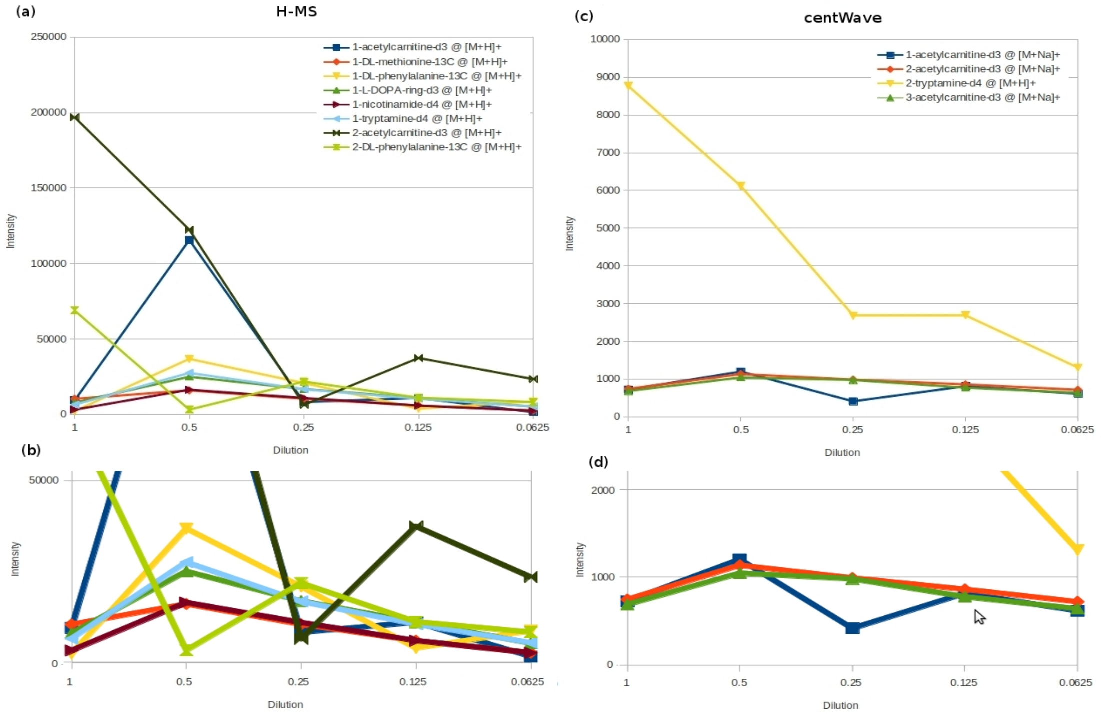


Figure 4.17: Intensity curves for all peaks detected by either H-MS, a) with closeup b), and centWave only, c) with closeup d), across the entire dilution series. The legend gives the molecule names, with the number in front indicating the replicate sample (1, 2 or 3). Lack of linear trend may be due to ion suppression or saturation effects.

relationship were inspected by eye to avoid falsely matched peaks. For example, Figure 4.18 shows three intensity plots for peaks that are detected and matched with a particular compound, but do not fit the linear function well. One possible explanation for the poor fit is that although the algorithm detects a peak, it may not be well quantified, e.g. the acetylcarnitine-d3+Na adduct is poorly characterised by centWave in Figure 4.18 b). These exceptions may also be due to unforeseen ion suppression effects or saturation in different samples.

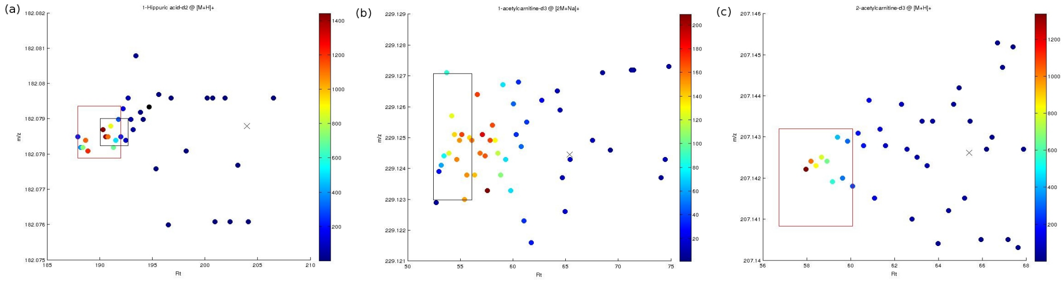


Figure 4.18: Raw intensity plots, with the colour scale indicating intensity (ion count) at a particular data point, of (a) Hippuric acid-d2+H peak detected by both H-MS and centWave in replicate 1, (b) acetylcarnitine-d3+Na peak detected by centWave in replicate 1 and (c) 2-acetylcarnitine-d3+H peak detected by H-MS in replicate 2. Peak output parameters shown for H-MS (red) and centWave (black) as boxes.

Rep	Peak	H-MS	centWave
1	Hippuric Acid-d2 @ [M+H]+	0.685	0.330
2	DL-Leucine-d3 @ [M+H]+	0.981	0.980
2	DL-Methionine-13C @ [M+H]+	0.925	0.911
2	Hippuric Acid-d2 @ [M+H]+	0.967	0.976
2	L-DOPA-Ring-d3 @ [M+H]+	0.971	0.892
2	Nicotinamide-d4 @ [M+H]+	0.976	0.962
3	Acetylcarnitine-d3 @ [M+H]+	0.994	0.997
3	DL-Leucine-d3 @ [M+H]+	0.993	0.990
3	DL-Methionine-13C @ [M+H]+	0.925	0.920
3	DL-Phenylalanine-13C @ [M+H]+	0.995	0.989
3	Hippuric Acid-d2 @ [M+H]+	0.911	0.930
3	L-DOPA-Ring-d3 @ [M+H]+	0.967	0.964
3	Nicotinamide-d4 @ [M+H]+	0.988	0.972

Table 4.9: R^2 for linear model fitted to dilution curves of peaks detected by both H-MS and centWave across all dilution levels.

Rep	Peak	R^2
1	Acetylcarnitine-d3 @ [M+H]+	0.037
1	DL-Methionine-13C @ [M+H]+	0.312
1	DL-Phenylalanine-13C @ [M+H]+	0.003
1	L-DOPA-Ring-d3 @ [M+H]+	0.004
1	Nicotinamide-d4 @ [M+H]+	0.000
1	Tryptamine-d4 @ [M+H]+	0.000
2	Acetylcarnitine-d3 @ [M+H]+	0.900
2	DL-Phenylalanine-13C @ [M+H]+	0.694

Table 4.10: R^2 for linear model fitted to dilution curves of peaks detected by H-MS but not centWave across all dilution levels.

Rep	Peak	R^2
1	Acetylcarnitine-d3 @ [M+Na]+	0.058
2	Acetylcarnitine-d3 @ [M+Na]+	0.002
2	Tryptamine-d4 @ [M+H]+	0.955
3	Acetylcarnitine-d3 @ [M+Na]+	0.000

Table 4.11: R^2 for linear model fitted to dilution curves of peaks detected by centWave but not H-MS across all dilution levels.

True Positive Rate

Peaks detected by H-MS in the SDS samples were also used to estimate the true positive detection rate. A random sample of 5% of all peaks detected across the fifteen samples for H-MS and centWave were independently scrutinised by 3 LC-MS experts within Imperial College London, UK, who gave their opinion on whether each detected peak was a true analyte peak or not, without knowledge of which algorithm had detected the peak. This gave us an estimated true positive discovery rate of 70.1% for both H-MS and centWave with standard deviations of 43% and 40% respectively. This survey again indicates that H-MS performs very similarly to centWave.

4.3.4 Conclusion

H-MS is a strategy that competes well alongside XCMS' centWave, as well as outperforming both the matchedFilter routine and the adaptive binning capabilities of apLC-MS. The m/z binning routine provides an effective alternative to fixed bin widths, whilst the Gaussian/EMG peak detection proves to be an efficient method with a sensitivity comparable with centWave.

Analysis of simulated ‘noise’ datasets showed apLCMS and matchedfilter to be more sensitive to false positives than centWave and H-MS when their parameters were relaxed. In exploring the synthetic urine it was found that apLCMS binning was error prone and both apLCMS and matchedfilter produced very imprecise peak parameters compared to centWave and H-MS. Some identified peaks illustrated how both centWave and H-MS sometimes did not optimally characterise the peaks either by including too many surrounding low intensity data-points (possible noise) or conversely did not capture some important high intensity points. This is likely due to both algorithms utilising some prescribed LC peak shape (mexican hat, Gaussian or EMG) which true analyte peaks will not always conform to.

The expert review of peaks detected for centWave and H-MS highlighted the similar performance with both algorithms estimated to have a good true positive rate of 70.1%. Analysis of the linearity of metabolite peak intensities detected across the standard dilution series also yielded very similar results but with both algorithms detecting some extra features that the other did not. H-MS detected several extra metabolite peaks consistently across dilutions despite detecting a smaller number of peaks overall, perhaps suggesting greater sensitivity and specificity over a range concentrations.

Evaluating and comparing performance of a peak detection algorithm effectively is difficult since even with synthetic mixtures we cannot be certain of the ‘correct answer’ because of the unknowns of metabolite interactions and effects. The gold standard for a peak detection algorithm really is undefined and this is particularly evident in the variance of responses to the expert review of detected peaks. Further rigorous testing could be performed by repeating this comparison across multiple instrumental configurations and examining the robustness of results in order to further illuminate strengths and weaknesses of the algorithm. For example, H-MS and centWave employ slightly different LC peak models that may be more suited to a particular type of LC columns or mobile phase flow method. For now, H-MS is shown to perform as well as one of the most commonly used methods in metabolomics and in some cases provides more accurate peak identification and description. Although it shares a similar error rate to centWave, H-MS seems to provide a slightly different view of the data to centWave and the algorithm may work well as a complementary/alternative routine in XCMS.

Chapter 5

Conclusion

According to a survey conducted by the American Society for Mass Spectrometry in 2009 (which can be found at `www.metabolomicssurvey.com`), a large number of researchers believe the biggest bottleneck within metabolomics is metabolite identification, followed by assigning biological significance then data processing/reduction. Although the reproducibility and accuracy of analytical techniques and the statistical methods for processing and interpreting the data have improved, the problems highlighted by the survey persist in the face of a huge diversity of molecular structures and variation of abundance [166]. Searching libraries of reference spectra to compare with experimental data is the most reliable method of metabolite identification but is a non-trivial task [167]. Metabolite spectral databases are continually expanding to include data for an increasing range of experimental conditions, biofluids and organisms. Attempting to match unknown spectral peaks with reference standards can throw up hundreds of candidate metabolites for both NMR and MS based techniques and there is a considerable amount of research devoted to developing spectral matching algorithms and ranking the results [167, 166, 53].

The more challenging alternative to using databases of reference spectra, is to interpret the metabolomic data in the context of metabolite molecular structure. Combining analytical techniques to give complementary information on the chemical structure of a molecule is essential in achieving the gold standard of metabolite identification. This method relies on the accuracy of the statistical methods employed to extract this information from complex spectral data not only by modelling data characteristics, but also the chemical processes occurring during data acquisition as a result of both the technique used and the interactions between metabolites and the contaminants introduced during sample processing.

The second major difficulty in metabolomics research, assigning biological significance, is a problem faced further along the typical experiment pipeline. Interpreting results within the systems biology framework can be a complex task, but once again database evolution may prove to be the way forward. In addition to the databases solely concerned with mapping metabolic pathways, metabolite databases such as the HMDB, increasingly provide cross-references to related metabolites, processes and organism responses and statuses. Work on integrating 'omics' data right across systems biology is ongoing and will undoubtedly be invaluable in placing metabolomics discoveries within a biologically meaningful context.

Thirdly, despite the obvious benefits, increasingly sensitive analytical techniques clearly pose a difficult challenge to metabolomics researchers. The vast datasets produced are incredibly difficult to explore in an intelligible way, so require pre-processing and dimension reduction. In Chapter 4 we reviewed a a number of different techniques common to LC-MS, of which many are applicable to other analytical platforms. This is a rich and diverse area of algorithm development benefitting immensely from the development of signal processing techniques in other disciplines, such as image processing. Similarly to the way in which metabolite identification relies on accurate data pre-processing methods, the advance of pre-processing algorithms relies on improved knowledge of biofluid composition and reliable models of interactions between the metabolites within it.

This thesis tackled three data analysis problems involved in widening the metabolomics bottleneck. In Chapter 2 we addressed the need to simulate metabolic data profiles in order to test and develop new statistical methods based on ^{1}H-NMR metabolomic profiles. MetAssimulo is a useful simulation tool for complex ^{1}H-NMR mixtures that is able to construct realistic human urine spectra with inter-metabolite correlations and peak positional variation. The user has complete control over numerous parameters involved in the simulation process, meaning

MetAssimulo has the capability to simulate profiles for many other biofluids and species. However, further work is needed to enable compatibility with other metabolite databases required for addressing other organisms. Development of techniques for modelling interactions and effects observed in ^{1}H-NMR spectra is ongoing, for example the effect of pH on peak positions examined in Chapter 3, so there will be opportunity to incorporate more realistic effects in the future.

Modelling phenomena affecting metabolite expression in ^{1}H-NMR spectra is a useful endeavour, not only for data simulators such as MetAssimulo, but also for the deconvolution of complex spectra and metabolite identification. The ability to predict peak positional variation dependent on the pH of a mixture would be invaluable in developing automated processing techniques. Conversely, identifying the pH of a biofluid from the position of a known metabolite peak or predicting the number of protonation sites of a molecule from titration curves would help in determining the chemical structure and identity of unknown metabolites. In Chapter.3 a Bayesian non-linear regression model was used to fit a polyprotic model and showed good parameter estimation. Attempts were also made to identify a method for choosing the model correctly reflecting the number of sites of the molecule. However, despite exploring several model selection criteria as well as modifying the model itself to include possible structured variation and borrow strength across multiple chemical shifts by employing random effects, no consistent solution was found. The model could benefit from being programmed using a reversible jump algorithm to perform model selection.

In Chapter 4, we turned our attention to the pre-processing of LC-MS data, in particular denoising and peak detection. After exploring some of the various methods employed to tackle the challenges faced in analysing LC-MS data, a novel Bayesian Partition Model (BPM) was developed with the aim of performing peak detection in the chromatographic domain. The BPM was effective at modelling changes in signal level, but ultimately proved too computationally intensive to compete with existing methods. A more efficient algorithm, H-MS was created using wavelet smoothing, adaptive binning and a combination of Gaussian and Exponentially Modified Gaussian peak detection. The resulting method was tested extensively using a synthetic urine sample to identify known compounds, with further testing performed on a standard dilution series to asses its ability to detect metabolites across a range of concentrations. In comparison to existing methods, H-MS outperformed both the adaptive package apLCMS and XCMS's matchedFilter. It competed well with XCMS's centWave method, often more appropriately describing peaks and proving more consistent across the dilution series. H-MS could become a useful package for those looking for an effective alternative to existing methods. Further work on decreasing the run-time of the algorithm could increase its efficiency and appeal to potential users. The algorithm could also be extended to perform additional pre-processing tasks, such as Retention Time alignment and matching peaks across samples.

Algorithms such as XCMS and H-MS could also benefit from the development of true 2-dimensional processing techniques, rather than performing peak detection on chromatograms after m/z binning. Methods capable of exploiting the information available as peaks drift in both dimensions is vital in avoiding the pitfalls of reducing analysis to a single dimension. Wavelet decomposition can be extended to two dimensions and has already been applied in analysing gas chromatogram differential mobility spectrometry signals with a similar 2-D structure to LC-MS [168] and used to de-noise LC-MS spectra in proteomics [169]. Given the level of complexity and size of digital imaging data across scientific disciplines, methods developed within the arena of image processing are likely to play an important role in the analysis of 2-D data for numerous areas of research. It is important for the metabolomic community to continue to mine these valuable interdisciplinary statistical resources in striving to transform noisy raw data into biologically meaningful information.

Metabolomics is a burgeoning field of study, with continuous technological advancement. Development of data modelling and simulation techniques incorporating evermore prior knowledge is vital to furthering our knowledge of metabolomics, the metabolome and its relationship to the other ‘omics’. The methods described within this thesis contribute to this development both in ^{1}H-NMR and LC-MS data analysis, in addition to providing further possible avenues of useful research.

Appendices

.1 Glossary of Abbreviations

AIC Akaike Information Criterion

apLCMS LC-MS processing software [66]

BIC Bayesian Information Criterion

DIC Deviance Information Criterion

CPO Conditional posterior ordinate

CWT Continuous wavelet transform

DWT Discrete wavelet transform

EIC Extracted Ion Chromatogram

EMG Exponentially modified gaussian model

ESI Electrospray Ionization

GC-MS Gas Chromatography Mass Spectrometry

HMDB Human Metabolome Database

H-MS LC-MS peak detection algorithm developed by the author

JAGS Just Another Gibbs Sampler

LC-MS Liquid Chromatography Mass Spectrometry

LPML Log Pseudo Marginal Likelihood

MCMC Markov Chain Monte Carlo

MSE Mean squared error

m/z Mass-to-charge ratio

NMR Nuclear Magnetic Resonance

NSSD NMR Standard Spectra Database

PCA Principal components analysis

PLS Partial least squares analysis

Rt Retention Time

SNR Signal to noise ratio

STOCSY Statistical Total Correlation Spectroscopy

TOF Time of Flight (mass spectrometry)

XCMS LC-MS and GC-MS processing software [46]

.2 Log Marginal Likelihood Estimation Algorithm

A linear iterative algorithm for estimating the log marginal likelihood of a model from [101]. Where L_t is a sample of the log likelihoods for $t = 1,..,T$. Firstly, the log likelihoods are mean-centred are exponentiated:

```
$for\ t = 1 : T$
    $f_t = exp(L_t - \bar{L_t})$
$end$
```

Where $\bar{L_t}$ is the mean of the log likelihood sample. The (centred) marginal likelihoods are then revised:

```
$\gamma_1 = 1$
$for\ j = 2 : 10$
    $A_j = 0$
    $B_j = 0$
    $for\ t = 1 : T$
        $A_j = A_j + \frac{f_t}{\delta\gamma_{j-1}+(1-\delta)f_t}$
        $B_j = B_j + \frac{1}{\delta\gamma_{j-1}+(1-\delta)f_t}$
```

```
end
$\gamma_j = \frac{(1-\delta)T+A_j}{(1-\delta)T/\gamma_{j-1}+B_j}$
end
```

where $\delta = 0.01$, giving the estimate:

$$log(\hat{ML}) = log(\gamma_{10}) + \bar{L}_t$$

.3 Marginal Likelihood Derivation for Bayesian Normal-Inverse-Gamma Models

This theorem is taken from [170] and provides a useful result for deriving the form of the marginal likelihood for Bayesian Normal-Inverse-Gamma models, used in the development of the Bayesian Partition Model earlier in the chapter.

Theorem

Suppose that the p-vector $\mathbf{y}$ and the n-vector θ are related via the conditional distribution

$$\mathbf{y} \mid \theta \sim N(F'\theta, V) \tag{1}$$

where the $(n \times p)$ matrix F and the $(p \times p)$ positive definite symmetric matrix V are constant. An equivalent statement is

$$\mathbf{y} = F'\theta + \nu \tag{2}$$

where $\nu \sim N(\mathbf{0}, V)$. The marginal distribution of θ is given by

$$\theta \sim N(\mathbf{a}, R) \tag{3}$$

where both $\mathbf{a}$ and R are constant, and that θ is independent of ν. Equivalently,

$$\theta = \mathbf{a} + \omega \tag{4}$$

where $\omega \sim N(\mathbf{0}, R)$ independently of ν. From these distributions it is possible to construct the joint distribution for $\mathbf{y}$ and the conditional for $(\theta \mid \mathbf{y})$.

Since $\theta = \mathbf{a}+\omega$ and $\mathbf{y} = F'\theta+\nu = F'\mathbf{a}+F'\omega+\nu$, then the vector $(\mathbf{y}', \theta')$ is a linear transformation of $(\nu', \omega')'$. By construction the latter has a multivariate normal distribution, so that $\mathbf{y}$ and θ are jointly normal. Furthermore:

1. $E[\theta] = \mathbf{a}$ and $V[\theta] = R$

2. $E[\mathbf{y}] = E[F'\theta = \nu] = F'E[\theta] + E[\nu] = F'\mathbf{a}$ and

$V[\mathbf{y}] = V[F'\theta + \nu] = F'V[\theta]F + V[\nu] = F'RF + V$

3. $C[\mathbf{y}, \theta] = C[F'\theta + \nu, \theta] = F'C[\theta, \theta] + C[\nu, \theta] = F'R$

It follows that

$$\begin{pmatrix} \mathbf{y} \\ \theta \end{pmatrix} \sim N \left(\begin{bmatrix} F'\mathbf{a} \\ \mathbf{a} \end{bmatrix}, \begin{bmatrix} F'RF + V & F'R \\ RF & R \end{bmatrix} \right) \tag{5}$$

Therefore, we have:

4. $\mathbf{y} \sim N[F'\mathbf{a}, F'RF + V]$

5. $(\theta \mid \mathbf{y}) \sim N(\mathbf{m}, C)$

where

$$\begin{aligned} \mathbf{m} &= \mathbf{a} + A\mathbf{e} \\ C &= R - AQA' \\ A &= RFQ^{-1} \\ Q &= F'RF + V \\ \mathbf{e} &= \mathbf{y} - F'\mathbf{a} \end{aligned}$$

.4 Code

The MATLAB code for MetAssimulo,the BPM and H-MS as well as the R code for the PH Model are included on disc.

Bibliography

[1] R. L. Last, A. D. Jones, and Y. Shachar-Hill, "Towards the plant metabolome and beyond," *Nat Rev Mol Cell Biol*, vol. 8, no. 2, pp. 167–174, 2007. 10.1038/nrm2098.

[2] M. Zhou, "Fernandez research group, georgia institute of technology website," 2010.

[3] K. C. Leptos, D. A. Sarracino, J. D. Jaffe, B. Krastins, and G. M. Church, "Mapquant: Open-source software for large-scale protein quantification," *Proteomics*, vol. 6, pp. 1770–1782, MAR 2006.

[4] B. K. Alsberg, A. M. Woodward, and D. B. Kell, "An introduction to wavelet transforms for chemometricians: A time-frequency approach," *Chemometrics And Intelligent Laboratory Systems*, vol. 37, pp. 215–239, JUN 1997.

[5] P. Du, W. A. Kibbe, and S. M. Lin, "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching," *Bioinformatics*, vol. 22, pp. 2059–2065, SEP 1 2006.

[6] J. K. Nicholson, J. C. Lindon, and E. Holmes, "'metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological nmr spectroscopic data," *Xenobiotica*, vol. 29, pp. 1181–1189, NOV 1999.

[7] 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*, vol. 22, pp. 245–252, MAY 2004.

[8] B. H. ter Kuile and H. V. Westerhoff, "Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway," *FEBS Letters*, vol. 500, pp. 169–171, JUL 6 2001.

[9] J. K. Nicholson, J. Connelly, J. C. Lindon, and E. Holmes, "Metabonomics: a platform for studying drug toxicity and gene function," *Nature Reviews Drug Discovery*, vol. 1, pp. 153–161, FEB 2002.

[10] S. M. Ibrahim and R. Gold, "Genomics, proteomics, metabolomics: what is in a word for multiple sclerosis?," *Current Opinion In Neurology*, vol. 18, pp. 231–235, JUN 2005.

[11] I. G. Khalil and C. Hill, "Systems biology for cancer," *Current Opinion In Oncology*, vol. 17, pp. 44–48, JAN 2005.

[12] J. K. Nicholson, "Global systems biology, personalized medicine and molecular epidemiology," *Molecular Systems Biology*, vol. 2, 2006.

[13] Y. Nikolsky, N. Nikolskaya, and A. Bugrim, "Biological networks and analysis of experimental data in drug discovery," *Drug Discovery Today*, vol. 10, pp. 653–662, MAY 1 2005.

[14] E. Holmes, R. L. Loo, J. Stamler, M. Bictash, I. K. S. Yap, Q. Chan, T. Ebbels, M. D. Iorio, I. J. Brown, K. A. Veselkov, M. L. Daviglus, H. Kesteloot, H. Ueshima, L. Zhao, J. K. Nicholson, and P. Elliott, "Human metabolic phenotype diversity and its association with diet and blood pressure," *Nature*, vol. 453, pp. 396–50, MAY 15 2008.

[15] K. Dettmer, P. A. Aronov, and B. D. Hammock, "Mass spectrometry-based metabolomics," *Mass Spectrometry Reviews*, vol. 26, pp. 51–78, JAN-FEB 2007.

[16] C. Chen, F. J. Gonzalez, and J. R. Idle, "Lc-ms-based metabolomics in drug metabolism," *Drug Metabolism Reviews*, vol. 39, no. 2-3, pp. 581–597, 2007.

[17] J. C. Lindon and J. K. Nicholson, "Analytical technologies for metabonomics and metabolomics, and multi-omic information recovery," *TRAC-Trends In Analytical Chemistry*, vol. 27, pp. 194–204, MAR 2008.

[18] V. Shulaev, "Metabolomics technology and bioinformatics," *Briefings In Bioinformatics*, vol. 7, pp. 128–139, JUN 2006.

[19] Z. Pan and D. Raftery, "Comparing and combining nmr spectroscopy and mass spectrometry in metabolomics," *Analytical And Bioanalytical Chemistry*, vol. 387, pp. 525–527, JAN 2007.

[20] S. Moco, R. J. Bino, R. C. H. D. Vos, and J. Vervoort, "Metabolomics technologies and metabolite identification," *TRAC-Trends In Analytical Chemistry*, vol. 26, pp. 855–866, OCT 2007.

[21] J. Forshed, H. Idborg, and S. P. Jacobsson, "Evaluation of different techniques for data fusion of lc/ms and h-1-nmr," *Chemometrics And Intelligent Laboratory Systems*, vol. 85, pp. 102–109, JAN 15 2007.

[22] W. S. Law, P. Y. Huang, E. S. Ong, C. N. Ong, S. F. Y. Li, K. K. Pasikanti, and E. C. Y. Chan, "Metabonomics investigation of human urine after ingestion of green tea with gas chromatography/mass spectrometry, liquid chromatography/mass spectrometry and h-1 nmr spectroscopy," *Rapid Communications In Mass Spectrometry*, vol. 22, pp. 2436–2446, AUG 2008.

[23] H. Y. Tong, D. E. Giblin, R. L. Lapp, S. J. Monson, and M. L. Gross, "Mass profile monitoring in trace analysis by gas-chromatography mass-spectrometry," *Analytical Chemistry*, vol. 63, pp. 1772–1780, SEP 1 1991.

[24] D. G. Robertson, M. D. Reily, R. E. Sigler, D. F. Wells, D. A. Paterson, and T. K. Braden, "Metabonomics: Evaluation of nuclear magnetic resonance (nmr) and pattern recognition technology for rapid in vivo screening of liver and kidney toxicants," *Toxilogical Sciences*, vol. 57, pp. 326–337, OCT 2000.

[25] S. Zhang, G. A. N. Gowda, T. Ye, and D. Raftery, "Advances in nmr-based biofluid analysis and metabolite profiling," *Analyst*, vol. 135, no. 7, pp. 1490–1498, 2010.

[26] P. Hore, *Nuclear Magnetic Resonance.* Oxford University Press Inc., 2004.

[27] J. C. Lindon, J. K. Nicholson, E. Holmes, and J. R. Everett, “Metabonomics: Metabolic processes studied by nmr spectroscopy of biofluids,” *Concepts In Magnetic Resonance*, vol. 12, no. 5, pp. 289–320, 2000.

[28] M. Defernez and I. Colquhoun, “Factors affecting the robustness of metabolite fingerprinting using h-1 nmr spectra,” *Phytochemistry*, vol. 62, pp. 1009–1017, MAR 2003.

[29] A. Craig, O. Cloareo, E. Holmes, J. Nicholson, and J. Lindon, “Scaling and normalization effects in nmr spectroscopic metabonomic data sets,” *Analytical Chemistry*, vol. 78, pp. 2262–2267, APR 1 2006.

[30] T. Ebbels and R. Cavill, “Bioinformatic methods in nmr-based metabolic profiling,” *Progress In Nuclear Magnetic Resonance Spectroscopy*, vol. 55, pp. 361–374, NOV 2009.

[31] O. Cloarec, M. E. Dumas, A. Craig, R. H. Barton, J. Trygg, J. Hudson, C. Blancher, D. Gauguier, J. C. Lindon, E. Holmes, and J. K. Nicholson, “Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic h-1 nmr data sets,” *Analytical Chemistry*, vol. 77, pp. 1282–1289, MAR 1 2005.

[32] D. Crockford, E. Holmes, J. Lindon, R. Plumb, S. Zirah, S. Bruce, P. Rainville, C. Stumpf, and J. Nicholson, “Statistical heterospectroscopy, an approach to the integrated analysis of nmr and uplc-ms data sets: Application in metabonomic toxicology studies,” *Analytical Chemistry*, vol. 78, pp. 363–371, JAN 15 2006.

[33] W. Gronwald, M. S. Klein, H. Kaspar, S. R. Fagerer, N. Nuernberger, K. Dettmer, T. Bertsch, and P. J. Oefner, “Urinary metabolite quantification employing 2d nmr spectroscopy,” *Analytical Chemistry*, vol. 80, pp. 9288–9297, DEC 1 2008.

[34] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. K. Wenger, H. Yao, and J. L. Markley, “Biomagresbank,” *Nucleic Acids Research*, vol. 36, pp. D402–D408, JAN 2008.

[35] D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M. A. Coutouly, I. Forsythe,

P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. E. Duggan, G. D. MacInnis, A. M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. D. Sykes, H. J. Vogel, and L. Querengesser, "Hmdb: the human metabolome database," *Nucleic Acids Research*, vol. 35, pp. –521, JAN 2007.

[36] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "Kegg for integration and interpretation of large-scale molecular data sets," *Nucleic Acids Research*, vol. 40, pp. D109–D114, JAN 2012.

[37] R. Caspi, T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, A. Pujar, A. G. Shearer, M. Travers, D. Weerasinghe, P. Zhang, and P. D. Karp, "The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases," *Nucleic Acids Research*, vol. 40, pp. D742–D753, JAN 2012.

[38] A. Kamburov, K. Pentchev, H. Galicka, C. Wierling, H. Lehrach, and R. Herwig, "Consensuspathdb: toward a more complete picture of cell biology," *Nucleic Acids Research*, vol. 39, pp. D712–D717, JAN 2011.

[39] T. Annesley, "Ion suppression in mass spectrometry," *Clinical Chemistry*, vol. 49, pp. 1041–1044, JUL 2003.

[40] J. Antignac, K. de Wasch, F. Monteau, H. De Brabander, F. Andre, and B. Le Bizec, "The ion suppression phenomenon in liquid chromatography-mass spectrometry and its consequences in the field of residue," *Analytica Chimica Acta*, vol. 529, pp. 129–136, JAN 24 2005. 5th EURORESIDUE Conference, Noordwijkerhout, NETheRLAndS, MAY 10-12, 2004.

[41] J. Zhang, E. Gonzalez, T. Hestilow, W. Haskins, and Y. Huang, "Review of peak detection algorithms in liquid-chromatography-mass spectrometry," *Current Genomics*, vol. 10, pp. 388–401, SEP 2009.

[42] R. E. Ardrey, *Liquid Chromatography - Mass Spectrometry: An Introduction.* John Wiley & Sons, Ltd, 2003.

[43] A. G. Marshall, C. L. Hendrickson, and G. S. Jackson, "Fourier transform ion cyclotron resonance mass spectrometry: A primer," *Mass Spectrometry Reviews*, vol. 17, pp. 1–35, JAN-FEB 1998.

[44] S. G. Villas-Boas, J. Nielsen, J. Smedsgaard, and M. Hansen, *Metabolome Analysis: An Introduction.* Wiley-Interscience, 2007.

[45] A. H. P. America and J. H. G. Cordewener, "Comparative lc-ms: A landscape of peaks and valleys," *Proteomics*, vol. 8, pp. 731–749, FEB 2008.

[46] C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak, "Xcms: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification," *Analytical Chemistry*, vol. 78, pp. 779–787, FEB 1 2006.

[47] M. J. Fredriksson, P. Petersson, J. K. Magnus, B. O. Axelsson, and D. Bylund, "An objective comparison of pre-processing methods for enhancement of liquid chromatography-mass spectrometry data," *Journal Of Chromatography A*, vol. 1172, pp. 135–150, NOV 23 2007.

[48] J. J. Baeza-Baeza and M. C. Garcia-Alvarez-Coque, "Prediction of peak shape as a function of retention in reversed-phase liquid chromatography," *Journal Of Chromatography A*, vol. 1022, pp. 17–24, JAN 2 2004.

[49] W. Lambert, "Pitfalls in lc-ms(-ms) analysis," *Toxichem und Krimtech*, vol. 71(2), no. 64, 2004.

[50] J. Listgarten and A. Emili, "Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry," *Molecular & Cellular Proteomics*, vol. 4, pp. 419–434, APR 2005.

[51] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda,

Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, and T. Nishioka, "Massbank: a public repository for sharing mass spectral data for life sciences," *Journal Of Mass Spectrometry*, vol. 45, pp. 703–714, JUL 2010.

[52] C. Smith, G. O'Maille, E. Want, C. Qin, S. Trauger, T. Brandon, D. Custodio, R. Abagyan, and G. Siuzdak, "Metlin - a metabolite mass spectral database," *Therapeutic Drug Monitoring*, vol. 27, pp. 747–751, DEC 2005. 9th International Congress of Therapeutic Drug Monitoring and Clinical Toxicology, Louisville, KY, APR 23-28, 2005.

[53] B. Zhou, J. F. Xiao, L. Tuli, and H. W. Ressom, "Lc-ms-based metabolomics," *Molecular Biosystems*, vol. 8, no. 2, pp. 470–481, 2012.

[54] B. S. Everitt, *The Cambridge Dictionary of Statistics.* Cambridge University Press, Cambridge, UK, 2nd ed., 2002.

[55] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis.* Texts in Statistical Science, Chapman & Hall/CRC, 2nd ed., 2004.

[56] W. A. Link and M. J. Eaton, "On thinning of chains in mcmc," *Methods in Ecology and Evolution*, vol. 3, no. 1, p. 112–115, 2012.

[57] M. K. Cowles and B. P. Carlin, "Markov chain monte carlo convergence diagnostics: A comparative review," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 883–904, 1996.

[58] D. Lunn, A. Thomas, N. Best, and D. Spiegelhalter, "Winbugs - a bayesian modelling framework: Concepts, structure, and extensibility," *Statistics And Computing*, vol. 10, pp. 325–337, OCT 2000.

[59] M. Plummer, "Jags: A program for analysis of bayesian graphical models using gibbs sampling," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 2003.

[60] J. Hao, W. Astle, M. De Iorio, and T. M. D. Ebbels, "Batman-an r package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a bayesian model," *Bioinformatics*, vol. 28, pp. 2088–2090, AUG 1 2012.

[61] Y. Sun, J. Zhang, U. Braga-Neto, and E. R. Dougherty, "Bpda2d-a 2d global optimization-based bayesian peptide detection algorithm for liquid chromatograph-mass spectrometry," *Bioinformatics*, vol. 28, pp. 564–572, FEB 15 2012.

[62] S. Chen, E. Deutsch, E. Yi, X. Li, D. Goodlettt, and R. Aebersold, "Improving mass and liquid chromatography based identification of proteins using bayesian scoring," *Journal Of Proteome Research*, vol. 4, pp. 2174–2184, NOV-DEC 2005.

[63] J. Jeong, X. Shi, X. Zhang, S. Kim, and C. Shen, "An empirical bayes model using a competition score for metabolite identification in gas chromatography mass spectrometry," *BMC Bioinformatics*, vol. 12, OCT 10 2011.

[64] J. Krumsiek, K. Suhre, T. Illig, J. Adamski, and F. J. Theis, "Bayesian independent component analysis recovers pathway signatures from blood metabolomics data," *Journal Of Proteome Research*, vol. 11, pp. 4120–4131, AUG 2012.

[65] H. Muncey, R. Jones, M. D. Iorio, and T. Ebbels, "Metassimulo: simulation of realistic nmr metabolic profiles," *BMC Bioinformatics*, vol. 11, no. 1, p. 496, 2010.

[66] T. Yu, Y. Park, J. M. Johnson, and D. P. Jones, "aplcms-adaptive processing of high-resolution lc/ms data," *Bioinformatics*, vol. 25, pp. 1930–1936, AUG 1 2009.

[67] M. L. Anthony, K. P. R. Gartland, C. R. Beddel, J. C. Lindon, and J. K. Nicholson, "Cephaloridine-induced nephrotoxicity in the fischer-344 rat – proton nmr spectroscopic studies of urine and plasma in relation to conventional clinical chemical and histopathological assessments of nephronal damage.," *Archives Of Toxicology*, vol. 66, pp. 525–537, OCT 1992.

[68] S. H. Moolenaar *et al.*, *Handbook of 1H-NMR Spectroscopy in Inborn Errors of Metabolism.* SPS Publications, Heilbronn, 2002.

[69] S. H. Moolenaar, U. F. H. Engelke, and R. A. Wevers, “Proton nuclear magnetic resonance spectroscopy of body fluids in the field of inborn errors of metabolism,” *Annals Of Clinical Biochemistry*, vol. 40, pp. 16–24, JAN 2003.

[70] O. Cloarec, M. E. Dumas, J. Trygg, A. Craig, R. H. Barton, J. C. Lindon, J. K. Nicholson, and E. Holmes, “Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in h-1 nmr spectroscopic metabonomic studies,” *Analytical Chemistry*, vol. 77, pp. 517–526, JAN 15 2005.

[71] P. E. Anderson, M. L. Raymer, B. J. Kelly, N. V. Reo, N. J. Delraso, and T. E. Doom, “Characterization of h-1 nmr spectroscopic data and the generation of synthetic validation sets,” *Bioinformatics*, vol. 25, pp. 2992–3000, NOV 15 2009.

[72] J. Hoch and A. Stern, *NMR Data Processing.* Wiley-Liss, Inc., London, 1996.

[73] C. Robert and G. Casella, *Monte Carlo Statistical Methods.* Springer-Verlag, New York, 1999.

[74] N. J. Higham, “Computing the nearest correlation matrix - a problem from finance,” *IMA Journal Of Numerical Analysis*, vol. 22, pp. 329–343, JUL 2002.

[75] A. Ross *et al.*, *NMR Spectroscopy Techniques for Application to Metabonomics. In: Lindon, J(ed) et al. The Handbook of Metabonomics and Metabolomics.* Elsevier, Oxford, 2007.

[76] Y. Chou, *Statistical Analysis.* Holt International, 1975.

[77] E. Bairaktari, K. Katopodis, K. C. Siamopoulos, and O. Tsolas, “Paraquat-induced renal injury studied by h-1 nuclear magnetic resonance spectroscopy of urine,” *Clinical Chemistry*, vol. 44, pp. 1256–1261, JUN 1998.

[78] M. Chadeau-Hyam, C. J. Hoggart, P. F. O’Reilly, J. C. Whittaker, M. D. Iorio, and D. J. Balding, “Fregene: Simulation of realistic sequence-level data in populations and ascertained samples,” *BMC Bioinformatics*, vol. 9, SEP 8 2008.

[79] S. M. Dudek, A. A. Motsinger, D. R. Velez, S. M. Williams, and M. D. Ritchie, "Data simulation software for whole-genome association and other studies in human genetics.," *Pacific Symposium on Biocomputing*, vol. 11, pp. 499–510, 2006.

[80] P. Gans, A. Sabatini, and A. Vacca, "Simultaneous calculation of equilibrium constants and standard formation enthalpies from calorimetric data for systems with multiple equilibria in solution," *Journal Of Solution Chemistry*, vol. 37, pp. 467–476, APR 2008. 17th Spanish-Italian Congress on the Thermodynamics of Metal Complexes/33rd Annual Congress of the Grupo-di-Termodinamica-die-Complessi, Seville, SPAIn, JUN 05-09, 2006.

[81] R. Allen, K. Box, J. Comer, C. Peake, and K. Tam, "Multiwavelength spectrophotometric determination of acid dissociation constants of ionizable drugs," *Journal Of Pharmaceutical And Biomedical Analysis*, vol. 17, pp. 699–712, AUG 1998.

[82] X. Xu and R. Hurtubise, "Determination of the pk(a) values of polycyclic aromatic hydrocarbon metabolites by capillary zone electrophoresis," *Journal Of Liquid Chromatography & Related Technologies*, vol. 22, no. 5, pp. 669–679, 1999.

[83] X. Kong, T. Zhou, Z. Liu, and R. C. Hider, "ph indicator titration: A novel fast pka determination method," *Journal Of Pharmaceutical Sciences*, vol. 96, pp. 2777–2783, OCT 2007.

[84] Z. Szakacs, M. Kraszni, and B. Noszal, "Determination of microscopic acid-base parameters from nmr-ph titrations," *Analytical And Bioanalytical Chemistry*, vol. 378, pp. 1428–1448, MAR 2004.

[85] C. Frassineti, S. Ghelli, P. Gans, A. Sabatini, M. S. Moruzzi, and A. Vacca, "Nuclear-magnetic-resonance as a tool for determining protonation constants of natural polyprotic bases in solution," *Analytical Biochemistry*, vol. 231, pp. 374–382, NOV 1 1995.

[86] A. Juffer, "Theoretical calculations of acid-dissociation constants of proteins," *Biochemistry And Cell Biology-Biochimie Et Biologie Cellulaire*, vol. 76, no. 2-3, pp. 198–209, 1998.

[87] H. Webb, B. M. Tynan-Connolly, G. M. Lee, D. Farrell, F. O'Meara, C. R. Sondergaard, K. Teilum, C. Hewage, L. P. McIntosh, and J. Nielsen, "Remeasuring hewl pk(a) values by nmr spectroscopy: Methods, analysis, accuracy, and implications for theoretical pk(a), calculations," *Proteins-Structure Function And Bioinformatics*, vol. 79, pp. 685–702, MAR 2011.

[88] J. Clayden, N. Greeves, S. Warren, and P. Worthers, *Organic Chemistry.* Oxford University Press Inc., 2001.

[89] J. L. Markley, "Observation of histidine residues in proteins by means of nuclear magnetic-resonance spectroscopy," *ACCOUNTS Of CHEMICAL Research*, vol. 8, no. 2, pp. 70–80, 1975.

[90] W. Schaller and A. D. Robertson, "Ph, ionic-strength, and temperature dependencies of ionization equilibria for the carboxyl groups in turkey ovomucoid 3rd domain," *Biochemistry*, vol. 34, pp. 4714–4723, APR 11 1995.

[91] G. Crisponi, V. NURCHI, T. PInTORI, and E. F. Trogu, "Computation of acidity constants of a polyprotic acid from nuclear-magnetic-resonance or uv-visible spectrophotometric data," *Analytica Chimica Acta*, vol. 184, pp. 77–85, JUN 30 1986.

[92] H. L. Surprenant, J. E. Sarneski, R. R. Key, J. T. Byrd, and C. N. Reilley, "C-13 nmr-studies of amino-acids - chemical-shifts, protonation shifts, microscopic protonation behavior," *Journal Of Magnetic Resonance*, vol. 40, no. 2, pp. 231–243, 1980.

[93] C. Frassineti, L. Alderighi, P. Gans, A. Sabatini, A. Vacca, and S. Ghelli, "Determination of protonation constants of some fluorinated polyamines by means of c-13 nmr data processed by the new computer program hypnmr2000. protonation sequence in polyamines," *Analytical and Bioanalytical Chemistry*, vol. 376, pp. 1041–1052, AUG 2003. 13th Spanish Italian Congress on the Thermodynamics of Metal Complexes/29th Annual Congress of Gruppo-di-Termodinamica-dei-Complessi, SANTIAGO COMPOSTE, SPAIn, JUN 02-06, 2002.

[94] R. I. Shrager, D. H. Sachs, A. N. Schechte, J. S. Cohen, and S. R. Heller, "Nuclear magnetic-resonance titration curves of histidine ring protons .2. mathematical models for interacting groups in nuclear magnetic-resonance titration curves," *Biochemistry*, vol. 11, no. 4, pp. 541–&, 1972.

[95] D. L. Rabenstein and T. L. Sayer, "Determination of microscopic acid dissociation-constants by nuclear magnetic-resonance spectrometry," *Analytical Chemistry*, vol. 48, no. 8, pp. 1141–1145, 1976.

[96] M. Noszal and Z. Szakacs, "Microscopic protonation equilibria of oxidized glutathione," *Journal Of Physical Chemistry B*, vol. 107, pp. 5074–5080, MAY 29 2003.

[97] A. Onufriev, D. A. Case, and G. M. Ullmann, "A novel view of ph titration in biomolecules," *Biochemistry*, vol. 40, pp. 3413–3419, MAR 27 2001.

[98] Z. Szakacs, G. Hagele, and R. Tyka, "H-1/p-31 nmr ph indicator series to eliminate the glass electrode in nmr spectroscopic pk(a) determinations," *Analytica Chimica Acta*, vol. 522, pp. 247–258, SEP 27 2004.

[99] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. AC-19(6), pp. 716–723, 1974.

[100] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6(2), pp. 461–464, 1978.

[101] P. Congdon, *Bayesian Statistical Modelling.* John Wiley & Sons, Ltd, 2006.

[102] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 64(4), pp. 583–639, 2002.

[103] S. Geisser and W. F. Eddy, "A predictive approach to model selection," *Journal of the American Statistical Association*, vol. 74, pp. 153–160, 1979.

[104] J. G. Ibrahim, M. Chen, and D. Sinha, *Bayesian Survival Analysis.* Springer, 2001.

[105] M. Chen, J. G. B. Ibrahim, and Q. Shao, "Power prior distributions for generalized linear models," *Journal of Statistical Planning and Inference*, vol. 84, pp. 121–137, 2000.

[106] M. A. Newton and A. E. Raftery, "Approximate bayesian inference with the weighted likelihood bootstrap," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 56(1), pp. 3–48, 1994.

[107] K. R.E. and R. A.E., "Bayes factors," *Journal of the American Statistical Association*, vol. 90, pp. 773–795, Jun 1995.

[108] K. R. Coombes, J. M. Koomen, K. A. Baggerly, R. Kobayashi, and J. S. Morris, "Understanding the characteristics of mass spectrometry data through the use of simulation," *Cancer Informatics*, vol. 1, pp. 41–52, 2005.

[109] C. C. Grigsby, M. R. Mateen, L. A. Tamburino, R. L. Pitsch, P. A. Shiyanov, and D. R. Cool, "Metabolite differentiation and discovery lab (meddl): A new tool for biomarker discovery and mass spectral visualization," *Analytical Chemistry*, vol. 88, no. 11, pp. 4386–4395, 2010.

[110] R. Sanchez-Ponce and F. P. Guengerich, "Untargeted analysis of mass spectrometry data for elucidation of metabolites and function of enzymes," *Analytical Chemistry*, vol. 79, pp. 3355–3362, MAY 1 2007.

[111] X. Kong and C. Reilly, "A bayesian approach to the alignment of mass spectra," *Bioinformatics*, vol. 25, pp. 3213–3220, DEC 15 2009.

[112] T. Skov, F. van den Berg, G. Tomasi, and R. Bro, "Automated alignment of chromatographic data," *Journal Of Chemometrics*, vol. 20, pp. 484–497, NOV-DEC 2006.

[113] C. Christin, H. C. J. Hoefsloot, A. K. Smilde, F. Suits, R. Bischoff, and P. L. Horvatovich, "Time alignment algorithms based on selected mass traces for complex lc-ms data," *Journal of Proteome Research*, vol. 9, pp. 1483–1495, MAR 2010.

[114] W. X. Wang, H. H. Zhou, H. Lin, S. Roy, T. A. Shaler, L. R. Hill, S. Norton, P. Kumar, M. Anderle, and C. H. Becker, "Quantification of proteins and metabolites by mass spec-

trometry without isotopic labeling or spiked standards," *Analytical Chemistry*, vol. 75, pp. 4818–4826, SEP 15 2003.

[115] H. Idborg, P. O. Edlund, and S. P. Jacobsson, "Multivariate approaches for efficient detection of potential metabolites from liquid chromatography/mass spectrometry data," *Rapid Communications In Mass Spectrometry*, vol. 18, no. 9, pp. 944–954, 2004.

[116] R. Matthiesen, "Methods, algorithms and tools in computational proteomics: A practical point of view," *Proteomics*, vol. 7, pp. 2815–2832, AUG 2007.

[117] P. Fryzlewicz, "Wavelet methods." To appear as an invited overview paper in Wiley Interdisciplinary Reviews: Computational Statistics. Available at http://stats.lse.ac.uk/fryzlewicz/articles.html., 2010.

[118] M. Misiti, Y. Misiti, G. Oppenheim, and J. Poggi, *Wavelets and their Applications*. Digital Signal & Image Processing Series, Wiley, 2010.

[119] C. A. Hastings, S. M. Norton, and S. Roy, "New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data," *Rapid Communications In Mass Spectrometry*, vol. 16, no. 5, pp. 462–467, 2002.

[120] M. Katajamaa and M. Oresic, "Data processing for mass spectrometry-based metabolomics," *Journal Of Chromatography A*, vol. 1158, pp. 318–328, JUL 27 2007.

[121] M. E. Monroe, N. Tolic, N. Jaitly, J. L. Shaw, J. N. Adkins, and R. D. Smith, "Viper: an advanced software package to support high-throughput lc-ms peptide identification," *Bioinformatics*, vol. 23, pp. 2021–2023, AUG 1 2007.

[122] D. M. Horn, R. A. Zubarev, and F. W. McLafferty, "Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules," *Journal Of The American Society For Mass Spectrometry*, vol. 11, pp. 320–332, APR 2000.

[123] M. Bellew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J. Eng, R. Fang, C. W. Lin, J. Chen, D. R. Goodlett, J. Whiteaker, A. Paulovich, and M. McIn-

tosh, "A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution lc-ms," *Bioinformatics*, vol. 22, pp. 1902–1909, AUG 1 2006.

[124] X. J. Li, E. C. Yi, C. J. Kemp, H. Zhang, and R. Aebersold, "A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry," *Molecular & Cellular Proteomics*, vol. 4, pp. 1328–1340, SEP 2005.

[125] S. M. Bozic, *Digital And Kalman Filtering : An introduction to discrete-time filtering and optimum linear estimation.* Edward Arnold, 2nd ed., 1994.

[126] C. A. Glasbey and G. W. Horgan, *Image Analysis for the Biological Sciences.* Statistics in Practise, John Wiley & Sons, Ltd, 1995.

[127] A. Banerji, "An introduction to image analysis using mathematical morphology," *IEEE Engineering in Medicine and Biology Magazine*, vol. 19, pp. 13–14, JUL-AUG 2000.

[128] A. W. Dowsey, J. A. English, F. Lisacek, J. S. Morris, G.-Z. Yang, and M. J. Dunn, "Image analysis tools and emerging algorithms for expression proteomics," *Proteomics*, vol. 10, no. 23, pp. 4226–4257, 2010.

[129] P. Du, R. Sudha, M. B. Prystowsky, and R. H. Angeletti, "Data reduction of isotope-resolved lc-ms spectra," *Bioinformatics*, vol. 23, pp. 1394–1400, JUN 1 2007.

[130] R. Tautenhahn, C. Bottcher, and S. Neumann, "Highly sensitive feature detection for high resolution lc/ms," *BMC Bioinformatics*, vol. 9, NOV 28 2008.

[131] S. Cappadona, F. Levander, M. Jansson, P. James, S. Cerutti, and L. Pattini, "Wavelet-based method for noise characterization and rejection in high-performance liquid chromatography coupled to mass spectrometry," *Analytical Chemistry*, vol. 80, no. 13, pp. 4960–4968, 2008.

[132] F. Mo, Q. Mo, Y. Chen, D. R. Goodlett, L. Hood, G. S. Omenn, S. Li, and B. Lin, "Waveletquant, an improved quantification software based on wavelet signal threshold

de-noising for labeled quantitative proteomic analysis.," *BMC Bioinformatics*, vol. 11, p. 219, 2010.

[133] P. Wang, P. Yang, J. Arthur, and J. Y. H. Yang, "A dynamic wavelet-based algorithm for pre-processing tandem mass spectrometry data.," *Bioinformatics*, vol. 26, no. 18, pp. 2242–9, 2010.

[134] Y. V. Karpievitch, E. G. Hill, J. S. Morris, K. R. Coombes, K. A. Baggerly, and J. S. Almeida, "Prepms, mass spectrometry graphical preprocessing tool," *Molecular & Cellular Proteomics*, vol. 5, p. 1311, OCT 2006.

[135] J. S. Morris, K. R. Coombes, J. Koomen, K. A. Baggerly, and R. Kobayashi, "Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum," *Bioinformatics*, vol. 21, pp. 1764–1775, MAY 1 2005.

[136] J. S. Morris, B. N. Clark, and H. B. Gutstein, "Pinnacle: a fast, automatic and accurate method for detecting and quantifying protein spots in 2-dimensional gel electrophoresis data," *Bioinformatics*, vol. 24, pp. 529–536, FEB 15 2008.

[137] K. M. Aberg, R. J. O. Torgrip, J. Kolmert, I. Schuppe-Koistinen, and J. Lindberg, "Feature detection and alignment of hyphenated chromatographic-mass spectrometric data - extraction of pure ion chromatograms using kalman tracking," *Journal Of Chromatography A*, vol. 1192, pp. 139–146, MAY 23 2008.

[138] G. K. Befekadu, M. G. Tadesse, and H. W. Ressom, "A bayesian based functional mixed-effects model for analysis of lc-ms data.," *Conf Proc IEEE Eng Med Biol Soc*, vol. 2009, pp. 6743–6746, 2009.

[139] J. S. Morris and R. J. Carroll, "Wavelet-based functional mixed models," *Journal Of The Royal Statistical Society Series B-Statistical Methodology*, vol. 68, no. Part 2, pp. 179–199, 2006.

[140] J. S. Morris, P. J. Brown, R. C. Herrick, K. A. Baggerly, and K. R. Coombes, "Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models," *Biometrics*, vol. 64, pp. 479–489, JUN 2008.

[141] A. V.-B. a. T. Pluskal, S.Castillo

[142] Y. Yasui, M. Pepe, M. L. Thompson, B. L. Adam, G. L. Wright, Y. S. Qu, J. D. Potter, M. Winget, M. Thornquist, and Z. D. Feng, "A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection," *Biostatistics*, vol. 4, pp. 449–463, JUL 2003.

[143] V. Andreev, T. Rejtar, H. S. Chen, E. V. Moskovets, A. R. Ivanov, and B. L. Karger, "A universal denoising and peak picking algorithm for lc-ms based on matched filtration in the chromatographic time domain," *Analytical Chemistry*, vol. 75, pp. 6314–6326, NOV 15 2003.

[144] R. Stolt, R. J. O. Torgrip, J. Lindberg, L. Csenki, J. Kolmert, I. Schuppe-Koistinen, and S. P. Jacobsson, "Second-order peak detection for multicomponent high-resolution lc/ms data," *Analytical Chemistry*, vol. 78, pp. 975–983, FEB 15 2006.

[145] Y. Wang, X. Zhou, H. Wang, K. Li, L. Yao, and S. T. C. Wong, "Reversible jump mcmc approach for peak identification for stroke seldi mass spectrometry using mixture model," *Bioinformatics*, vol. 24, pp. –407, JUL 1 2008. 16th ISMB Conference on Intelligent Systems for Molecular Biology, Toronto, CANADA, JUL 19-23, 2008.

[146] A. Cruz-marcelo, R. Guerra, M. Vannucci, Y. T. Li, C. C. Lau, and T. K. Man, "Comparison of algorithms for pre-processing of seldi-tof mass spectrometry data," *Bioinformatics*, vol. 24, no. 19, pp. 2129–2136, 2008.

[147] C. Yang, Z. Y. He, and W. C. Yu, "Comparison of public peak detection algorithms for maldi mass spectrometry data analysis," *BMC Bioinformatics*, vol. 10, p. 4, 2009.

[148] S. Chen, M. Li, D. Hong, D. Billheimer, H. Li, B. J. Xu, and Y. Shyr, "A novel comprehensive wave-form ms data processing method," *Bioinformatics*, vol. 25, no. 6, pp. 808–814, 2009.

[149] R. A. Davis, A. J. Charlton, J. Godward, S. A. Jones, M. Harrison, and J. C. Wilson, "Adaptive binning: An improved binning method for metabolomics data using the undec-

imated wavelet transform," *Chemometrics And Intelligent Laboratory Systems*, vol. 85, pp. 144–154, JAN 15 2007.

[150] P. Carvalho, G. Rocha, and M. P. Hobson, "A fast bayesian approach to discrete object detection in astronomical data sets - powellsnakes i," *Monthly Notices Of The Royal Astronomical Society*, vol. 393, pp. 681–702, MAR 1 2009.

[151] S. C. Dass, "Assessing fingerprint individuality in presence of noisy minutiae," *IEEE Transactions on Information Forensics and Security*, vol. 5, pp. 62–70, MAR 2010.

[152] R. L. Carter and B. J. N. Blight, "A bayesian change-point problem with an application to the prediction and detection of ovulation in women," *Biometrics*, vol. 37, no. 4, pp. 743–751, 1981.

[153] J. A. Hartigan, "Partition models," *Communications In Statistics-Theory And Methods*, vol. 19, no. 8, pp. 2745–2756, 1990.

[154] D. Barry and J. A. Hartigan, "Product partition models for change point problems," *Annals of Statistics*, vol. 20, pp. 260–279, MAR 1992.

[155] D. Barry and J. A. Hartigan, "A bayesian-analysis for change point problems," *Journal of the American Statistical Association*, vol. 88, pp. 309–319, MAR 1993.

[156] B. P. Carlin, A. E. Gelfand, and A. F. M. Smith, "Hierarchical bayesian-analysis of changepoint problems," *Applied Statistics-Journal Of The Royal Statistical Society Series C*, vol. 41, no. 2, pp. 389–405, 1992.

[157] J. Chen and A. K. Gupta, "On change point detection and estimation," *Communications In Statistics-Simulation And Computation*, vol. 30, no. 3, pp. 665–697, 2001.

[158] S. Chib, "Estimation and comparison of multiple change-point models," *Journal of Econometrics*, vol. 86, pp. 221–241, OCT 1998.

[159] E. Moreno, G. Casella, and A. Garcia-Ferrer, "An objective bayesian analysis of the change point problem," *Stochastic Environmental Research And Risk Assessment*, vol. 19, pp. 191–204, AUG 2005.

[160] D. A. Stephens, “Bayesian retrospective multiple-changepoint identification,” *Applied Statistics-Journal Of The Royal Statistical Society Series C*, vol. 43, no. 1, pp. 159–178, 1994.

[161] A. O’Hagan, *Bayesian Inference*, vol. 2B of *Kendall’s Advanced Theory of Statistics.* Halsted Press, 1994.

[162] D. G. T. Denison, C. C. Holmes, B. K. Mallick, and A. F. M. Smith, *Bayesian Methods for Nonlinear Classification and Regression.* PROBABILITY & STATISTICS, John Wiley & Sons, Ltd, 2002.

[163] D. Kwon, M. Vannucci, J. J. Song, J. Jeong, and R. M. Pfeiffer, “A novel wavelet-based thresholding method for the pre-processing of mass spectrometry data that accounts for heterogeneous noise.,” *Proteomics*, vol. 8, no. 15, pp. 3019–29, 2008.

[164] V. Di Marco and G. Bombi, “Mathematical functions for the representation of chromatographic peaks,” *Journal Of Chromatography A*, vol. 931, pp. 1–30, OCT 5 2001.

[165] O. Schulz-Trieglaff, N. Pfeifer, C. Gröpl, O. Kohlbacher, and K. Reinert, “Lc-mssim - a simulation software for liquid chromatography mass spectrometry data,” *BMC Bioinformatics*, vol. 9, OCT 8 2008.

[166] J. F. Xiao, B. Zhou, and H. W. Ressom, “Metabolite identification and quantitation in lc-ms/ms-based metabolomics,” *TRAC-Trends In Analytical Chemistry*, vol. 32, pp. 1–14, FEB 2012.

[167] S. Neumann and S. Boecker, “Computational mass spectrometry for metabolomics: Identification of metabolites and small molecules,” *Analytical And Bioanalytical Chemistry*, vol. 398, pp. 2779–2788, DEC 2010.

[168] W. Zhao, S. Sankaran, A. M. Ibanez, A. M. Dandekar, and C. E. Davis, “Two-dimensional wavelet analysis based classification of gas chromatogram differential mobility spectrometry signals,” *Analytica Chimica Acta*, vol. 647, pp. 46–53, AUG 4 2009.

[169] D. C. Compton and R. R. Snapp, “Detecting trace components in liquid chromatography mass spectrometry data sets with two-dimensional wavelets,” in *Wavelet Applications In Industrial Processing V* (F. Truchetet and O. Laligant, eds.), vol. 6763 of *Proceedings Of The Society Of Photo-Optical Instrumentation Engineers (SPIE)*, (1000 20TH ST, PO BOX 10, BELLInGHAM, WA 98227-0010 USA), SPIE, SPIE-InT SOC OPTICAL ENGInEERInG, 2007. Conference on Wavelet Applications in Industrial Processing V, Boston, MA, SEP 11-12, 2007.

[170] M. West and J. Harrison, *Bayesian Forecasting and Dynamic Models*. Springer Series in Statistics, Springer, 2nd ed., 1997.