Bayesian Methods for Metabolomics

Lifeng Ye

supervised by

Prof. Maria De Iorio (First Supervisor)

Dr. Alexandros Beskos (Second Supervisor)

Department of Statistical Science

University College London

September 4, 2020

I, Lifeng Ye, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Metabolomics, the large-scale study of small molecules, enables the underlying biochemical activity and state of cells or tissues to be directly captured. Nuclear Magnetic Resonance (NMR) Spectroscopy is one of the major data capturing techniques for metabolomics, as it provides highly reproducible, quantitative information on a wide variety of metabolites. This work presents possible solutions for three problems involved to aid the development of better algorithms for NMR data analysis. After reviewing relevant concepts and literature, we first utilise observed NMR chemical shift titration data for a range of urinary metabolites and develop a theoretical model of chemical shift using a Bayesian statistical framework and model selection procedures to estimate the number of protonation sites, a key parameter to model the relationship between chemical shift variation and pH and usually unknown in uncatalogued metabolites. Secondly, with the aim of obtaining explicit concentration estimates for metabolites from NMR spectra, we discuss a Monte Carlo Co-ordinate Ascent Variational Inference (MC-CAVI) algorithm that combines Markov chain Monte Carlo (MCMC) methods with Co-ordinate Ascent VI (CAVI), demonstrate MC-CAVI's suitability for models with *hard constraints* and compare MC-CAVI's performance with that of MCMC in an important complex model used in NMR spectroscopy data analysis. The third distribution seeks to improve metabolite identification, one of the biggest bottlenecks in metabolomics and severely hindered by resonance overlapping in one-dimensional NMR spectroscopy. In particular, we present a novel Bayesian method for widely used two-dimensional

(2D) ^{1}H *J*-resolved (JRES) NMR spectroscopy, which has considerable potential to accurately identify and quantify metabolites within complex biological samples, through combining B-spline tight wavelet frames with theoretical templates. We then demonstrate the effectiveness of our approach via analyses of JRES datasets from serum and urine.

Impact Statement

Metabolomics, as an expression of genetic and environmental factors, is essential to facilitating our further comprehension of how humans and other organisms function as individuals and interacting complex systems. ^{1}H Nuclear Magnetic Resonance (NMR) spectroscopy is one of the main techniques used for metabolite data acquisition but usually large and heavily structured. The aim of this thesis is to design pioneering analytical approaches and statistical methods to aid the development of better algorithms for NMR data analysis. The establishment of our work can have a great contribution inside academia. First, the estimation of the number of protonation sites from NMR spectroscopic data may be valuable for the future development of algorithms for analysis of metabolomic ^{1}H NMR spectra including alignment, annotation and peak fitting. Second, our discuss about the efficacy of MC-CAVI algorithm helps researchers in choosing algorithms for NMR data analysis and deciding future directions for algorithms improvement. Third, our Bayesian model for JRES NMR spectroscopy data analysis, which has the capacity to incorporate information from previous experiments and reduce resonance overlapping, benefits teams who focus on deconvolution and quantification of metabolites (e.g. Metabolomics Research Group in RIKEN Center for Sustainable Resource Science) and gives a guidance for developing Bayesian models for other spectroscopic datasets.

The work of the thesis can be effortlessly extended outside academia. Pharmaceutical industry has studied metabolomics for nearly three decades now. Research

group with these companies can utilise our model for JRES to their advantage in accurate metabolite deconvolution and quantification, which will assist a wide range of applications such as genetically modified plants development, disease-screening and drug toxicity and pharmacology study. Our approach for estimating the number of protonation sites will also give assistance to the development of NMR data analysing packages. In addition, the discussion of MC-CAVI algorithm can be applied to a variety of areas including the booming field of artificial intelligence where computational efficiency is especially important due to the extremely large datasets.

Acknowledgements

I would like to thank Prof. Maria De Iorio and Dr. Alexandros Beskos for their invaluable and professional guidance during my PhD research. I am very grateful to Dr. Timothy Ebbels, Imperial College London, for his patient help in understanding metabolomics and NMR data. I would also like to thank Dr. Andreas Heinecke, Yale-NUS College, for his help in learning spline wavelets and implementing JRES in Chapter 4. I would like to acknowledge Dr. William Astle, Cambridge University, Dr. Jie Hao, Shanghai Jiao Tong University, and Dr. Jianliang Gao, Imperial College London, for their great help in implementing new BATMAN algorithms.

I also thank Dr. Gregory Tredwell, Australian National University, and Dr. Jake Bundy, Imperial College London, for providing the titration series data used in Chapter 2. Dr. Goncalo Miguel Gomes Da Graca, Imperial College London, is acknowledged for providing the serum data and urine data used in the testing of JRES model in Chapter 4. My financial support is acknowledged from the CSC-UCL Joint Research Scholarship.

I really appreciate Dr. Andreas Heinecke and Sandy for the help during my visit to Singapore. I am very grateful to my family and my friends especially Cyro, Marco, Wei, Yiren and Ze for their love, support and encouragement throughout my academic career, without whom I would never get this far. Last but not least, I thank my dog Biku for always being there.

Contents

1 Introductory Material **22**

1.1 Metabolomics 22

1.2 ^{1}H Nuclear Magnetic Resonance Spectroscopy 23

1.3 JRES 29

1.4 Bayesian Inference 31

1.5 Variational Inference 38

1.6 A Bayesian Model of NMR Spectra 43

1.6.1 Prior Specification 45

1.6.2 MCMC Algorithm 47

1.7 Aims 48

2 Bayesian Estimation of the Number of Protonation Sites from NMR Spectroscopic Data **50**

2.1 Background 50

2.2 Methods 51

2.2.1 The model 51

2.2.2 Specification of Prior Knowledge 53

2.2.3 Prior Specification for pKa 55

2.2.4 Data 55

2.3 Results and Discussion 56

2.3.1 Metabolites with incorrectly estimated number of protonation sites 61
2.4 Conclusions 63
2.5 Compliance with Ethical Standards 64

3 On the Efficacy of Monte Carlo Implementation of CAVI 65
3.1 Background 65
3.2 MC-CAVI Algorithm 66
3.2.1 Description of the Algorithm 66
3.2.2 Applicability of MC-CAVI 68
3.2.3 Theoretical Justification of MC-CAVI 69
3.2.4 Stopping Criterion and Sample Size 72
3.3 Numerical Examples – Simulation Study 74
3.3.1 Simulated Example 1 74
3.3.2 Variance Reduction for BBVI 77
3.3.3 Simulated Example 2: Model with Hard Constraints 79
3.4 Application to ^{1}H NMR Spectroscopy 86
3.5 Conclusion 90

4 Bayesian deconvolution and quantification of metabolites from *J*-resolved NMR spectroscopy 92
4.1 Background 92
4.2 Modelling 95
4.2.1 Modelling of catalogued metabolite signal 97
4.2.2 Modelling of uncatalogued metabolite signal 100
4.2.3 Likelihood 106
4.3 Prior specifications 107
4.4 MCMC algorithm 115
4.5 Simulation study 117

4.6 Performance on urine and serum spectra 119
4.6.1 Jres spectra 119
4.6.2 Comparison between 1D NMR and 2D JRES deconvolution and quantification, and with bucketing 122
4.6.3 1D NMR spectra and comparison with BATMAN 124
4.7 Conclusion 128

5 Final Remarks **131**
5.1 General conclusion 131
5.2 Future research 134

Appendices **137**
A1 JAGS Code 137
A2 Proof of Lemma 1 139
A3 Proof of Theorem 1 140
A4 Gradient Expressions for BBVI 141
A5 MC-CAVI Implementation of BATMAN 142
A6 Details of MCMC strategy 145
A7 Simulation study details 148
A8 Additional figures and convergence from urine data 150
A9 Serum metabolite quantification from JRES spectrum 156
A10 Sensitivity analysis 160
A11 Posterior distributions for serum and urine spectra 165

Bibliography **171**

List of Figures

1.1 Example of a NMR multiplet signal resulting from spin-spin coupling. 27

1.2 Example of ^{1}H NMR Spectra 28

1.3 Example of a JRES spectrum surface plot. The x-axis corresponds to the chemical shift and is measured in parts per million (ppm) of the resonant frequency of a standard peak. The y-axis corresponds to the J-coupling information and shows the distance of each peak from the center of the resonance measured in Hz/F. The standardized intensity on the z-axis is proportional to the concentration of the corresponding metabolite. 31

1.4 Left panel: a trace plot of a well-mixed Markov chain; Right panel: a trace plot of a poorly-mixed Markov chain 36

2.1 Upper panel: ^{1}H NMR spectra with pH adjusted from 2 to 12. Lower left panel: Observed chemical shift positions (y) of 51 resonances. Lower right panel: Fitted chemical shift positions ($\tilde{\delta}$) for the 51 resonances. Only resonances with correct q predicted are shown. 57

2.2 Measured Chemical Shift Changes for Acetate, Alanine, Threonine and TTMethylHistidine with literature pKa Values (yellow vertical line), fitted pKa values (green vertical line) and the fit of the theoretical model (red line). The x-axis corresponds to pH and y-axis corresponds to ppm. 61

2.3 Examples of resonances with incorrectly estimated numbers of sites: Taurine, Citrate, Creatinine, Imidazole with literature pKa Values (yellow vertical line), fitted pKa values (green vertical line) and the fit of the theoretical model (red line). The x-axis corresponds to pH and y-axis corresponds to ppm. 62

3.1 Tracplots of ζ (left), θ (right) from application of CAVI in Simulated Example 1. 75

3.2 Traceplot of $\widehat{\mathbb{E}}(\tau)$ generated by MC-CAVI for Simulated Example 1, using $N = 10$ for the first 10 iterations of the algorithm, and $N = 10^3$ for the rest. The y-axis gives the values of $\widehat{\mathbb{E}}(\tau)$ across iterations. 76

3.3 Traceplot of $\widehat{\mathbb{E}}(\tau)$ under different settings of A-B-C (respectively, the value of N in the burn-in period, the number of burn-in iterations and the value of N after burn-in) for Simulated Example 1. 76

3.4 Plot of convergence time versus variance of $\widehat{\mathbb{E}}(\tau)$ (left panel) and versus Monte Carlo sample size N (right panel). 77

3.5 Model fit (left panel), traceplots of ϑ (middle panel) and traceplots of θ (right panel) for the three algorithms: MCMC (first row), MC-CAVI (second row) and BBVI (third row) – for Example Model 2 – when allowed 100secs of execution. In the plots showing model fit, the green line represents the data without noise, the orange line the data with noise; the blue line shows the corresponding posterior means and the grey area the pointwise 95% posterior credible intervals. 85

3.6 Density plots for the true posterior of ϑ (blue line) – obtained via an expensive MCMC – and the corresponding approximate distribution provided by MC-CAVI. 85

3.7 Traceplots of Parameter Value against Number of Iterations after the burn-in period for β_3 (upper left panel), β_4 (upper right panel), β_9 (lower left panel) and $\delta_{4,1}$ (lower right panel). The y-axis corresponds to the obtained parameter values (the mean of the distribution q for MC-CAVI and traceplots for MCMC). The red line shows the results from MC-CAVI and the blue line from MCMC. Both algorithms are executed for the same (approximately) amount of time. 88

3.8 Comparison of MC-CAVI and MCMC in terms of Spectral Fit. The upper panel shows the Spectral Fit from MC-CAVI algorithm. The lower panel shows the Spectral Fit from MCMC algorithm. The x-axis corresponds to chemical shift measure in ppm. The y-axis corresponds to standard density. 89

3.9 Comparison of Metabolites Fit obtained with MC-CAVI and MCMC. The x-axis corresponds to chemical shift measure in ppm. The y-axis corresponds to standard density. The upper left panel shows areas around ppm value 2.14 (β_4 and β_9). The upper right panel shows areas around ppm 2.66 (β_6). The lower left panel shows areas around ppm value 3.78 (β_3 and β_9). The lower right panel shows areas around ppm 7.53 (β_{10}). 90

4.1 Peak configurations of some common multiplet types. The x-axis indicates chemical shift while the y-axis indicates the J-coupling. The upper panel shows a doublet with chemical shift $\delta^{\star}_{mu}$ and peak offset ζ_{muv}. The lower panel shows a triplet and quadruplet. 98

4.2 Effect of additional local shrinkage applied to framelet coefficients of selected targeted regions. For ease of visualization, spectra are vectorised columnwise and plotted in 2D. On the x-axis we report the chemical shift region of the multiplet, on the y-axis their intensities. The top panel shows the templates of the metabolites Valine and Isoleucine that are targeted. The theoretical template of the multiplet structure of Valine is doublet-doublet-doublet with proton intensity ratio 3:3:1 (recall that we do not include one of the Valine multiplets in the analysis), while that of Isoleucine is triplet-doublet-doublet with proton intensity ratio 3:3:1. Additional local shrinkage is applied in the experiment shown in the bottom panel to the regions of high proton multiplets, i.e. to the first three columns in the lower panel, meaning that estimation is driven by Valine. Compared to the middle panel, in which no additional local shrinkage is applied, this strategy leads to improved accuracy of the concentration estimation for the metabolites. 113

4.3 Top panel: Comparison between the logarithm of the true relative concentrations and the estimated relative concentrations obtained with our method on the ten mixtures. Bottom panel: Performance comparison between our approach and the bucketing method on the ten simulated biological mixtures. 117

4.4 Deconvolution surface plot from urine JRES spectrum for the region around 1.337ppm, where the resonance is generated by Lactate. For ease of visualization we plot the fit on a grid of equally spaced points with distance 0.002ppm for the chemical shift axis. . . 120

4.5 Posterior relative concentration estimates and posterior standard deviations using our method on the urine spectrum from 1D NMR measurements, from 2D JRES measurements as well as using the bucketing method on the 2D JRES measurements. Valine is chosen as baseline. For four of the five targeted metabolites the posterior means of the estimates obtained using the second dimension differ by more than 25%. The figure shows 95% credible intervals. Note that bucketing only produces point estimates with no quantification of uncertainty of the estimate. 123

4.6 Deconvolution of selected regions from the urine 1D NMR data. The x-axis corresponds chemical shift in ppm and y-axis to intensities. The top panel shows resonances generated by Valine, Leucine, Isoleucine and 3-Hydrixybutyrate. The lower middle panel and lower right panel show resonances generated by Alanine and Lactate, respectively. The lower left panel shows resonances generated by untargeted metabolites and weak signals from Valine and Isoleucine. 125

4.7 Deconvolution of resonances generated by untargeted metabolites for a selected region from a urine 1D NMR spectrum. The *x*-axis corresponds to chemical shift in ppm and the *y*-axis to the intensities. The measured spectrum is shown in black, while the B-spline frame component of our model is plotted in red and the Symlet 6 wavelet component of BATMAN in blue. 127

4.8 Heat maps for intensities from the urine JRES dataset. The *x*-axis corresponds to chemical shift in ppm, the *y*-axis to *J*-coupling in MHz/F. Plots show original data (upper panel), overall fitting, i.e., metabolite and framelet fitting (middle panel), and fitting of metabolites only (lower panel). Multiplets in the lower panel from left to right: Valine (3.601ppm), Alanine, Lactate, 3-Hydroxybutyrate, Valine (1.029ppm), Valine (0.976ppm), Leucine (0.95ppm), Leucine (0.94ppm). The Isoleucine fit is not visible in the lower panel as its concentration estimation is close to zero. . . . 130

1 Heatmap of the intensities of the ten simulated biological mixtures (from 1 to 10 row-wise). The *x*-axis corresponds to chemical shift in ppm, *y*-axis to *J*-coupling in MHz/F. 148

2 Heatmap of the simulated JRES baseline spectrum with bin boundaries (red). The *x*-axis corresponds to chemical shift in ppm, *y*-axis to *J*-coupling in MHz/F. 149

3 Deconvolution of selected regions from the urine JRES data. Panels show resonances generated by Valine (top panel) and Leucine (bottom panel). On the x-axis we report the chemical region of the multiplet. On the y-axis we report the intensity of the multiplet. The data is vectorised columnwise and plotted in 2D. Original data is displayed in black, untargeted component of the model is displayed in red. 151

4 Deconvolution of selected regions from the urine JRES spectrum. Top panel shows resonances generated by Valine and Isoleucine. The latter is not present at a detectable level. Bottom panel shows resonances generated by Alanine, Lactate and 3-Hydroxybutyrate. On the x-axis we report the chemical shift region of the multiplet. On the y-axis we report the intensity of the multiplet. The data is vectorised columnwise and plotted in 2D. 152

5 Traceplot of the log-likelihood. 153

6 Traceplot of $\log \beta$ (concentration parameter) of Valine (right panel) and 3-Hydrixybutyrate (left panel). The x-axis corresponds to the number of iterations, the y-axis corresponds to the logarithm of the sample value. 153

7 Traceplots of $\log \theta_{ijl}$ for four randomly chosen framelet parameters. The x-axis corresponds to the number of iterations, the y-axis corresponds to the logarithm of sample values. 154

8 Traceplot of $\log \lambda$ (precision parameter). The x-axis corresponds to the number of iterations, the y-axis corresponds to the logarithm of sample value 154

9 Traceplot of $\log \delta$ (chemical shift parameter) of Valine 1.029ppm (left panel) and Leucine 0.95ppm (right panel). The x-axis corresponds to the number of iterations, the y-axis corresponds to the logarithm of sample values. 155

10 Traceplot of $\log \zeta$ (J-coupling parameter) of Isoleucine 3.65ppm (left panel) and Alanine (right panel). The x-axis corresponds to the number of iterations, the y-axis corresponds to the logarithm of sample values. 155

11 Surface plot of deconvolution for region around 0.958ppm from the serum JRES spectrum. In this region the resonance is generated by the second multiplet of Leucine. For ease of visualization we plot the fit on a ppm-grid of 0.002 equally spaced points. 157

12 Heat maps for intensities from serum JRES spectrum. The x-axis corresponds to chemical shift in ppm, y-axis to J-coupling in MHz/F. Upper panel shows the original data, middle panel shows the overall fitting (metabolite fitting and wavelet fitting), and lower panel shows metabolite fitting only. The multiplets in the lower panel from left to right are: Isoleucine (3.66ppm), Valine (3.60ppm), Valine (1.03ppm), Isoleucine (1.00ppm), Valine (0.98ppm), Leucine (0.95ppm), Leucine (0.94ppm), Isoleucine (0.93ppm). 158

13 Deconvolution of selected regions from serum JRES spectrum. Panels show resonances generated by Valine (top), Leucine (middle) and Isoleucine (bottom). On the x-axis we report the chemical shift region of the multiplet. On the y-axis we report the intensity of the multiplet. The data is vectorised columnwise and plotted in 2D. . . 159

14 Comparison of posterior means of the shift in peak locations obtained with different prior distributions for the scalar precision parameter λ. 162

15 Comparison of posterior means of concentration parameters obtained with different prior distributions for the scalar precision parameter λ. 162

16 Comparison of shift in peak locations with different prior distributions for the global shrinkage parameter τ. 163

17 Comparison of estimated concentration with different prior distributions for the global shrinkage parameter τ. 163

18 Comparison of posterior estimates of shift in peak locations obtained with different prior distributions for the local shrinkage parameters μ_{ijl}. 164

19 Comparison of posterior estimates of concentration with different prior distributions for the local shrinkage parameters μ_{ijl}. 164

20 Posterior distributions of concentration parameters of the serum spectra. 165

21 Posterior distributions of chemical shift parameters of the serum spectra. 166

22 Posterior distribution of the J-coupling parameters of the serum spectra. 167

23 Posterior distributions of the concentration parameters of the urine spectra. 168

24 Posterior distributions of the chemical shift parameters of the urine spectra. 169

25 Posterior distribution of the J-coupling parameters of the urine spectra. 170

List of Tables

2.1 Details about Parameters 54

2.2 Comparison of the literature number of sites and the number estimated by the model 56

2.3 Probability of different numbers of protonation sites, estimated number of protonation sites and literature number of protonation sites for 51 resonances from 32 metabolites in human urine. Rows with correctly estimated numbers of sites are shown in bold. 58

2.4 Literature and Modelled Results of Acetate, Alanine, Threonine and TTMethylHistidine 60

3.1 Results of MC-CAVI for Simulated Example 1. 76

3.2 Summary of results: last two rows show the average for the corresponding parameter (in horizontal direction) and algorithm (in vertical direction), after burn-in (the number in brackets is the corresponding standard deviation). All algorithms were executed for 10^2secs. The first row gives some algorithmic details. 84

3.3 Estimation of β obtained with MC-CAVI and MCMC. (The coefficients of β for which the posterior means obtained with the two algorithms differ by more than 1.0e-4 are shown in bold.) 88

4.1 Details about Parameters 115

4.2 Comparison of modelling strategy between BATMAN and our approach. 124

4.3 Comparison of effective sample sizes (ESS) and integrated autocorrelation times (IAC) of the coefficients of the uncatalogued signal component between BATMAN and our method. We report summary statistics of the ESS and IAC values of all wavelet/framelet coefficients. 127

1 True relative concentrations (RC), posterior estimates of relative concentrations obtained with our model (BAYES) and estimates obtained by bucketing/binning (BIN) for ten simulated biological mixtures of Valine (Val), Isoleucine (Iso), Threonine (Thr) and Glucose (Glu). 149

2 Posterior relative concentration estimates (RC) and posterior standard deviations (SD) using our method on urine spectra from 1D NMR measurements as compared to 2D JRES measurements and to 2D bucketing/binning. Note that no standard deviation is available for the bucketing/binning method. Valine is chosen as baseline. For four of the five targeted metabolites the posterior means of the estimates obtained using the second dimension differ by more than 25%. 150

Chapter 1

Introductory Material

1.1 Metabolomics

In the post-genomic era, utilising "omic" techniques to investigate different levels of biological organisation is gaining increasing and extensive popularity in both industry and academia. Being different from other "omic" measures, the small molecules, also known as *metabolites*, within cells, biofluids, tissues or organisms, and their concentrations directly reflect the underlying biochemical activities and states of cells or tissues. Therefore, when explaining the relationship between genes and the overall function of a system, *metabolomics*, i.e. the large-scale study of *metabolome* (the complete set of metabolites) and their interaction within an organism, more closely reveals the activities of the organism at a functional level [56].

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 generic modification" [95]. Although the terms "metabolomics", "metabonomics", metabolic "fingerprinting" or "profiling" were assigned subtly different definitions originally, they are usually interchangeably used nowadays.

Generally, around 2,000 major metabolites for humans are considered. This number, however, increases substantially when secondary metabolites from bacteria, fungi, or plants are considered [27, 35].

Focusing on high-throughput identification and quantification of metabolites [96], metabolomics brings an extra dimension to our knowledge of biological systems because metabolic fluxes, i.e. the rate of turnover of molecules through a metabolic pathway, are regulated not only by gene expression, but also by additional factors, such as the abundance of metabolites as substrates (molecules acted upon by enzymes) [122]. Metabolites in biofluids are in dynamic equilibrium with those metabolites in cells and tissues so that their metabolic profile reflects the state transition of an organism caused by environmental or disease factors. Therefore, as an expression of genetic and environmental factors, metabolomics is essential to facilitate our further comprehension of how humans and other organisms function as individuals and interacting complex systems.

Metabolomics is utilised extensively from studying drug toxicity and pharmacology to disease-screening for conditions, such as, cancer or diabetes [65, 69, 75, 94, 97]. For example, "personalised health-care solutions", which is the ultimate customisation of healthcare, requires metabolomics for quick medical diagnosis to identify disease. Besides, in agriculture, metabolomics can help us to develop genetically modified plants and to estimate associated risks of difference modification by obtaining a glimpse of their complex biochemistry through informative snapshots acquired at different time points during plant growth.

1.2 ^{1}H Nuclear Magnetic Resonance Spectroscopy

Almost all experiments in metabolomics require identification or quantification of metabolites in complex biological mixtures, usually biofluids or tissue samples.

Thus, depending on the aims or priorities of a study or experiment, a variety of data-capturing techniques is employed in metabolomics, each with its own advantages and disadvantages. Among these techniques, ^{1}H Nuclear Magnetic Resonance (NMR) spectroscopy is one of the main techniques used for metabolite data acquisition [80, 115] because of its many advantages:

- NMR spectroscopy requires minimal sample preparation so that the analysis process is highly reproducible.
- NMR spectroscopy is able to give an almost global metabolite profile including structural information, which enables the identification of the most abundant metabolites.
- NMR spectroscopy has the potential to detect nearly all proton-containing metabolites and allows metabolites to be detected simultaneously without preselection.
- NMR spectroscopy is capable of measuring concentrations as low as 100μM [111] and even lower with some techniques such as cryoprobe technology [141].

With all these advantages, ^{1}H NMR is widely applied and research in NMR based metabolomics has obtained substantial attention in biomedical sciences, with numerous applications in the areas of biology and medicine, including biochemistry [104, 99], oncology [57, 64], disease diagnostics [18, 11], epidemiology [66, 131], genetics [70, 34], organism classification [20, 88], and toxicology [81, 59]. For instance, [11] show that in patients affected by head and neck squamous cell carcinoma and undergoing radio-/chemo-radiotherapy real-time dynamic changes in the serum metabolome can be detected at the beginning of the treatment using NMR-based metabolomics. This metabolic alterations are characteristic for malnutrition or cachexia and their early detection enables identifying and monitoring patients

with a higher risk of weight loss.

NMR spectroscopy, however, is not flawless. One major drawback of NMR spectroscopy, compared with other analytical methods (e.g. mass spectrometry), is its relatively poor sensitivity: there are usually thousands of metabolites in biofluids, but a typical NMR spectrum only contains signals from a few hundred of the most abundant metabolites.

NMR spectroscopy is based on the fact that in a static 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 is dependent upon the strength of interaction between the magnetic moment of the nucleus and the field [68]. This energy difference is chemical shift, which can be measured by using electromagnetic radiation of a certain frequency which drives the nuclei to shift between states.

The position and number of chemical shifts can be used to diagnose the chemical structure of the molecule. 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 relies on the local magnetic field experienced by each nucleus. The local magnetic field is dependent upon the extent to which molecular orbitals shield the influence of the external spectrometer field. Therefore the chemical shift reveals the bonding configuration and chemical structure of the metabolite. While Hz is the fundamental fre-

quency unit of NMR, the frequency observed is based on the strength of the magnetic field. Thus the position of each peak is given in a scale of parts per million (ppm) instead through dividing the response frequency (in Hz) by the carrier frequency (in MHz) [68] and the y-axis is measured and standardised by dividing peak heights to the peak heights of an internal standard (e.g. 3-(Trimethylsilyl)-Propionic acid-D4 (TSP)).

2. Spin-spin coupling (also known as J-coupling or scalar coupling) is the phenomenon of magnetic interactions between close nuclei. Due to spin-spin coupling, a proton has more than one resonant frequency resulting in a juxtaposition of peaks named as a "multiplet" (as shown in Figure 1.1), whose pattern is determined by the chemical structure of the molecule.

3. For a given metabolite, with the assumption that there are no differential relaxation effects, integrated peak area is proportional to the number of existing ^{1}H nuclei and allows quantification of the concentration of the metabolite.

A typical NMR spectrum is characterised by a 1-dimensional (1D) signal, which consists of a series of resonance intensity measurements taken over a grid of frequencies (a scale of ppm of an internal standard), where the x-axis corresponds to the resonant frequency (usually plotted to decrease from left to right), which is controlled artificially through experiments. The y-axis corresponds to the resonance intensity, which are the observations of the experiment. The intensity varies in proportion to metabolite concentrations detected and metabolites are represented by "peaks" in the spectral data, where the instrument has registered the presence of a molecular species within the biofluid. Another important component of a NMR spectrum is the chemical shift, which is the resonant frequency (location on the x-axis) of a peak and hence is inferred from the observations. An example of ^{1}H NMR spectra is shown in Figure 1.2.

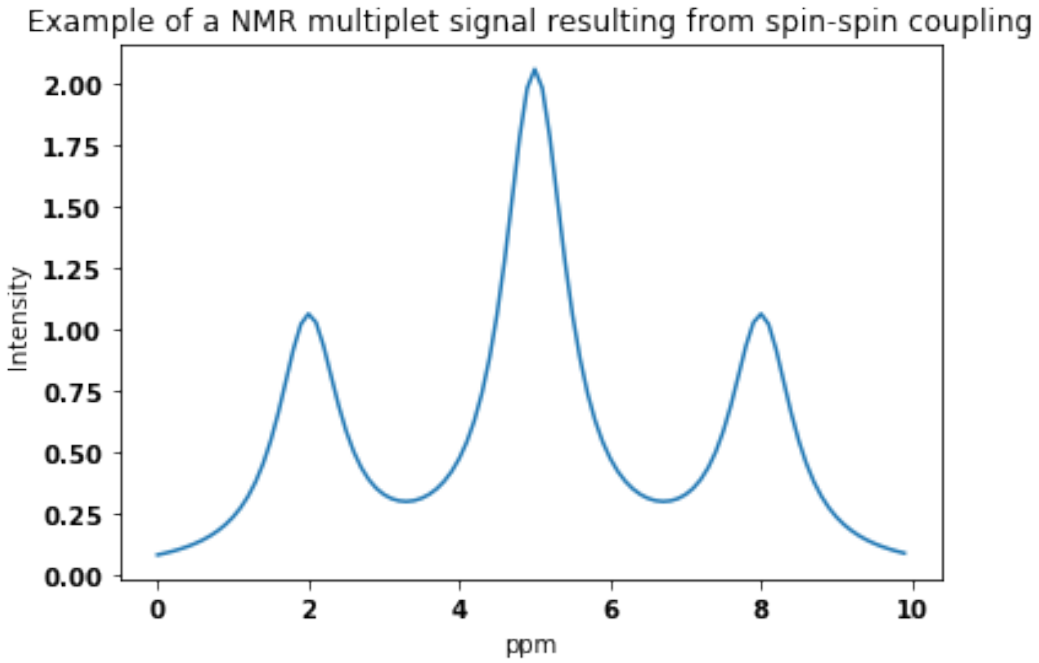

Figure 1.1: Example of a NMR multiplet signal resulting from spin-spin coupling.

The spectrum from a pure compound will comprise a "signature" of peaks, which contains some information of the structure of the compound. Under ideal conditions, each peak has the form of a Lorentzian curve. A zero-centred, standardized Lorentzian function can be represented by the following equation (i.e., the pdf of a Cauchy distribution with scale parameter $\gamma/2$):

$$\ell_\gamma(x) = \frac{2\gamma}{\pi(4x^2+\gamma^2)}, \tag{1.1}$$

where γ is interpreted as the "peak-width at half-height" (or "linewidth") and x is the resonance frequency. The NMR spectrum of a complex mixture can be effectively approximated by a linear combination of several spectra from pure compounds, i.e a biofluid spectrum containing K different metabolites can be treated as K-dimensional object, in which each dimension is the concentration signal of a single metabolite [82]. This superposition of peaks and multiplets generates a

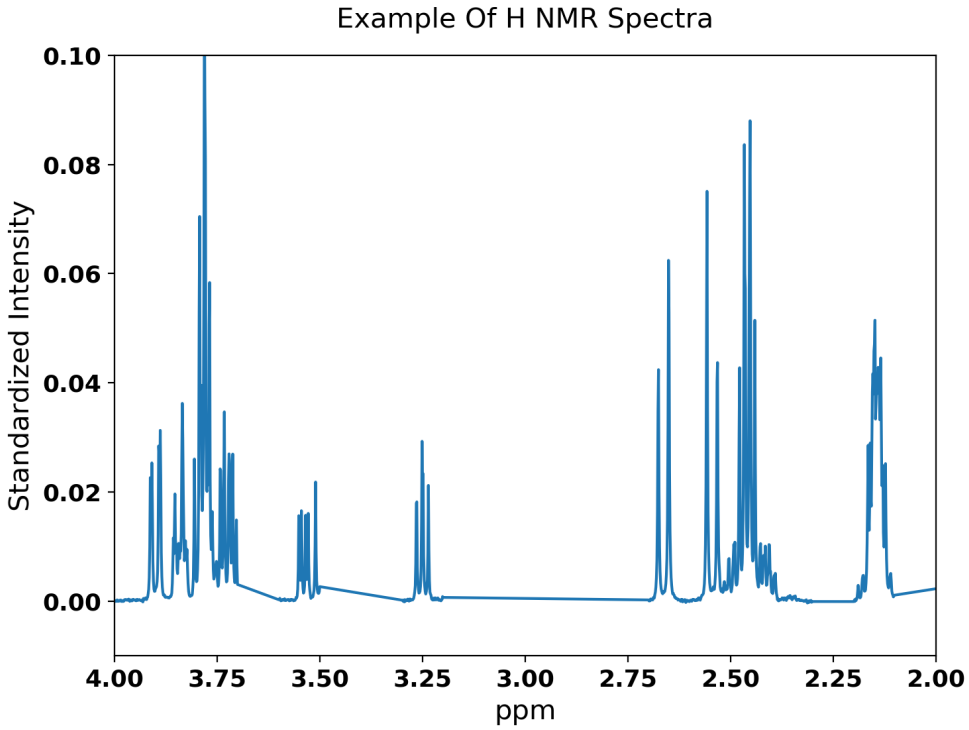

Figure 1.2: Example of ^{1}H NMR Spectra

complex spectrum where individual signature patterns overlap. These spectra can be further complicated by the shift of peak position, which is the mixed consequences of matrix effects, variation in experimental conditions and differences in the chemical properties of the sample, e.g pH value and the strength of other ionic species in the mixture [33]. These problems, combined with background noise and the presence of contaminants, result in difficulty in designing automated algorithms for deconvolution and quantification of metabolites from NMR spectroscopy.

Identification and quantification of metabolite signals in NMR data has made huge progresses as a consequence of the development of several databases aimed to document metabolite data, such as Biological Magnetic Resonance Bank (BMRB) [129], which provides the information on molecules including peptides, proteins, and nucleic acids, and the Human Metabolome Database (HMDB) [136, 137, 138, 139],

which focuses more specifically on small molecule metabolites discovered in the human body. Although these databases are valuable, there are still several challenges in curating the "ideal" data resources. One major problem is that these databases are inherently limited in their coverage. The other problem is that the experimental conditions may be inconsistent for different metabolite data. For instance, there might be much variation in pH in different experiments.

Astle et al. [3] developed a Bayesian model, which incorporated information available in online databases on the patterns of spectral resonance generated by human metabolites, to automate peak assignment and spectral deconvolution for 1D ^{1}H NMR spectra in the frequency domain. This model and its specially designed MCMC strategy are implemented in the R package BATMAN [60]. However, this model cannot fully address the problems of target signals being overlapped by other sharp signals, which are not explicitly modelled. This problem is particularly pronounced in crowded spectral regions. Therefore, it is of paramount importance to develop appropriate statistical approaches to precisely identify and quantify metabolites within complex biological samples, so that the capability of metabolomics can be fully realised.

1.3 JRES

The full capability of metabolomics cannot be achieved until appropriate approaches are established to precisely identify and quantify metabolites within complex biological samples. Spectral deconvolution and identification can be substantially improved by going from one- to two-dimensional spectra at the expense of prolonged experimental time. Two-dimensional (2D) NMR methods have considerable potential and become increasingly popular in metabolomics. Compared to 1D spectra, peak overlap in 2D spectra is greatly diminished because spin magnetization is transferred between different nuclear spins and provides more so-

phisticated spectra. The introduction of an additional dimension allows for a better representation of metabolites, which greatly aids biomarker identification.

A popular 2D method for metabolomics is 2D ^{1}H J-resolved NMR spectroscopy (JRES), first introduced by Aue et al. [5]. Unlike other 2D methods, such as correlation spectroscopy (COSY) [4, 17] or total correlation spectroscopy (TOCSY) [32], which use J-coupling to correlate chemical shifts of the coupling spins, JRES disperses the overlapping resonances into a second dimension and provides a metabolic fingerprint in a relatively short acquisition time because of the low number of increments recorded in the spin-spin coupling dimension. In 1D spectra, much of the peak overlap is due to each resonance being split into multiple peaks by J-coupling. Moving this dispersion into a separate dimension in JRES therefore significantly reduces congestion, and enhances metabolite identification and estimation [84].

2D JRES spectra are collections of convolved peaks, of which Figure 1.3 shows an example. Each spectral peak corresponds to magnetic nuclei resonating in the biological mixture represented by a pair of frequency coordinates determining the displacement of the peak in the (x,y)-plane. The x-axis corresponds to the chemical shift and is measured in parts per million (ppm) of the resonant frequency of a standard peak. The y-axis corresponds to the J-coupling information and shows the distance of each peak from the centre of the resonance in Hz/F, where F is the operating frequency of the spectrometer in MHz. Volume under each peak on the z-axis is proportional to the concentration of the corresponding metabolite in the biological mixture.

Resonance frequencies of magnetic nuclei are largely determined by their molecular environment, that is, the chemical structure of the molecules in which they are embedded and the configuration of their chemical bonds within the molecules. Consequently, every metabolite has a characteristic molecular 2D ^{1}H J-resolved

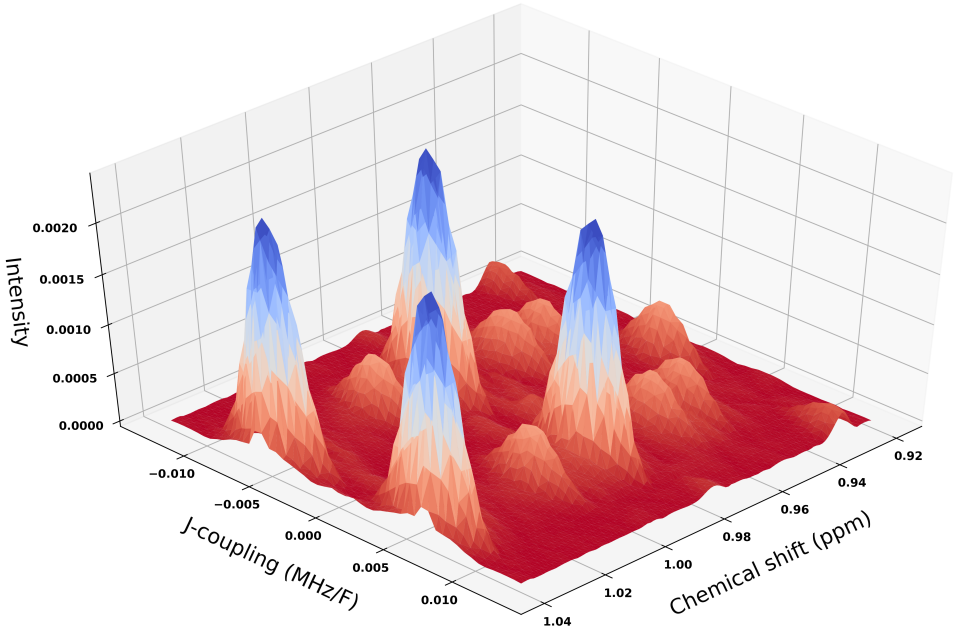

Figure 1.3: Example of a JRES spectrum surface plot. The x-axis corresponds to the chemical shift and is measured in parts per million (ppm) of the resonant frequency of a standard peak. The y-axis corresponds to the J-coupling information and shows the distance of each peak from the center of the resonance measured in Hz/F. The standardized intensity on the z-axis is proportional to the concentration of the corresponding metabolite.

NMR signature, i.e. presents itself as a convolution of peaks that appear in specific positions in the 2D JRES spectrum. The peaks of a signature often have significantly different chemical shifts and J-coupling information, and so appear widely separated in a spectrum.

1.4 Bayesian Inference

In order to perform inference of both 1D and 2D NMR data, which is large and heavily structured, sophisticated statistical techniques are necessary. Bayesian methodology has the capacity of being able to incorporate expert knowledge, database information and previous experiment results into the prior distribution, and is extensively applied in the field of metabolomics for several purposes (e.g., latent variable analysis, network/pathway analysis, variable selection/dimension

reduction and spectral deconvolution).

Bayesian inference is an approach to statistics in which all forms of uncertainty are expressed with probability. To start with, we need to specify a model, which we assume is adequate, to describe the situation of interest. Then, based on our beliefs about the situation before seeing the data $\mathbf{y}$, we formulate a prior distribution $p(\theta|\boldsymbol{\phi})$ over θ, where θ denotes the unknown parameter of the model and $\boldsymbol{\phi}$ denotes the hyperparameters. After observing the data, we apply Bayes' theorem (alternatively Bayes' law or Bayes' rule) to obtain a posterior distribution for the unknown parameters, combining the information from both the prior and the data. This process can be stated mathematically as the following equation:

$$p(\theta|\mathbf{y},\boldsymbol{\phi}) = \frac{p(\mathbf{y}|\theta)p(\theta|\boldsymbol{\phi})}{p(\mathbf{y}|\boldsymbol{\phi})}, \tag{1.2}$$

where $p(\theta|\mathbf{y},\boldsymbol{\phi})$ is called the posterior probability of the parameter θ. From the posterior distribution of the unknown parameter ($p(\theta|\mathbf{y},\boldsymbol{\phi})$), we are able to compute predictive distributions for future observations and describe the data generating process through, for example, calculating the expectation and variance of the unknown parameter (θ). $p(\mathbf{y}|\theta)$ is the likelihood function of the data and $p(\mathbf{y}|\boldsymbol{\phi})$ is the marginal likelihood, which can be calculated by the following equation:

$$p(\mathbf{y}|\boldsymbol{\phi}) = \int_{\theta} p(\mathbf{y}|\theta)p(\theta|\boldsymbol{\phi})d\theta. \tag{1.3}$$

$p(\mathbf{y}|\boldsymbol{\phi})$ is also referred to as "model evidence".

In the case of multi-parameter models, where $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$, deriving the probability distribution of the parameters of interest, say θ_1, requires averaging over the remaining parameters. The marginal distribution $p(\theta_1|\mathbf{y})$ needs to be derived from

the joint posterior distribution $p(\boldsymbol{\theta}|y) = p(\theta_1, \theta_2, ..., \theta_k|y)$ by:

$$p(\theta_1|y) = \int_{\theta_k} \int_{\theta_{k-1}} ... \int_{\theta_2} p(\theta_1, \theta_2, ..., \theta_k|y) d\theta_2 d\theta_3 ... d\theta_k \tag{1.4}$$

or alternatively

$$p(\theta_1|y) = \int_{\theta_k} \int_{\theta_{k-1}} ... \int_{\theta_2} p(\theta_1|\theta_2, ..., \theta_k, y) p(\theta_2, \theta_3, ..., \theta_k|y) d\theta_2 d\theta_3 ... d\theta_k. \tag{1.5}$$

These high dimensional posterior distributions are usually rather challenging to calculate either analytically or numerically. The problem of making inference on this type of distributions can be addressed by employing Markov Chain Monte Carlo (MCMC) methods. MCMC methods are a class of algorithms for sampling (simulation of random draws) from a complex probability distribution, say $f(x)$ by constructing a Markov chain that has $f(x)$ as its equilibrium distribution. A discrete-time Markov chain is a stochastic process $\{X_t : t = 0, 1, 2, ...\}$ satisfying the Markov property (also known as the memoryless property) such that:

$$P(X_n = x_n|X_0 = x_0, X_1 = x_1, ..., X_{n-1} = x_{n-1}) = P(X_n = x_n|X_{n-1} = x_{n-1}) \tag{1.6}$$

i.e. the probability distribution of future state is only determined by the present state, not on the entire observation history [44]. For a continuous-time Markov chain $\{X_t : t \geq 0\}$ with state space S, the Markov property follows:

$$P(X(t) = j|X(s) = i, X(t_{n-1}) = i_{n-1}, \ldots, X(t_1) = i_1) = P(X(t) = j|X(s) = i), \tag{1.7}$$

where $0 \leq t_1 \leq t_2 \leq \cdots \leq t_{n-1} \leq s \leq t$ is any non-decreasing sequence of $n+1$ times and $i_1, i_2, \ldots, i_{n-1}, i, j \in S$ are any $n+1$ states in the state space, for any integer $n \geq 1$. Once a large enough sample has been obtained, all the essential features of the probability distribution of interest can be approximated and summarized to

any degree of accuracy.

There are many MCMC methods, such as the Metropolis-Hastings algorithm, which is simple but practical and can be used, in principle, to obtain random samples from any arbitrarily complicated target distribution of any dimension that is known up to a normalizing constant. Suppose we need to draw samples from $f(x)$, a complex probability distribution whose normalization constant is impossible or extremely difficult to calculate. The Metropolis-Hastings algorithm requires only the value of a function, say $g(x)$, which is proportional to $f(x)$:

$$g(x) \propto c * f(x) \tag{1.8}$$

where c is the normalizing constant. For each iteration t, a candidate value Y is proposed for X_{t+1} using a "proposal distribution", say $q(x)$, with acceptance probability $\alpha(X_t, Y)$, which has the following form:

$$\alpha(X_t, Y) = \min\{1, \frac{g(Y)q(X_t|Y)}{g(X_t)q(Y|X_t)}\}. \tag{1.9}$$

A special case of the Metropolis-Hastings algorithm is the Gibbs sampling (or Gibbs sampler), whose probability of acceptance is equal to one. Gibbs sampling is applicable when the conditional distribution of each variable is known and is easy (or at least, easier) to sample from while the joint distribution is not known explicitly or is difficult to sample from directly. Suppose a large sample of $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)$ needs to be obtained from the joint distribution $f(\theta_1, \theta_2, ..., \theta_k)$. Then for each draw ($t = 1, 2, ...$), $\theta_i^{(t)}$ is sampled from the conditional distribution $p(\theta_i^{(t)}|\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_{i-1}^{(t)}, \theta_{i+1}^{(t)}, ..., \theta_k^{(t)})$, which is proportional to the joint distribution $f(\theta_1, \theta_2, ..., \theta_k)$. In other words, each variable is sampled given the most recently updated values of rest variables. There are various ways to extend

Metropolis-Hastings algorithm, for example, the Metropolis-adjusted Langevin algorithm (MALA), based on solving the Langevin diffusion,

$$dX_t = \frac{1}{2}\nabla \log f(X_t)dt + dB_t, \tag{1.10}$$

where B_t is the standard Brownian motion and ∇f denotes the gradient of f. Although computing the gradient at each iteration requires extra time, there is strong evidence that the MALA algorithm provides noticeable speed-ups in convergence [109]. Particle Markov chain Monte Carlo (PMCMC) is another extension of the Metropolis-Hastings algorithm, where sequential Monte Carlo (SMC) is introduced to design efficient high dimensional proposal distributions. With appropriate choice of proposal distribution, PMCMC is suitable for sampling from a target distribution with much high dimension and complex patterns of dependence [2]. The application of adaptive rejection sampling (ARS) methods for sampling from the full-conditional densities [90] is another extension of the Metropolis-Hastings algorithms.

Burn-in is the practice of throwing away some iterations at the beginning of an MCMC run. In theory, the Markov chain eventually converges to the desired distribution, but it is possible that the initial samples follow a very different distribution, especially when the chosen starting point is from a region with low density. Therefore, a burn-in period is often utilised so that the effect of initial values on posterior inference is minimized since it is unlikely to start with a good initial point. In practice, we usually choose a large value, say M, and assume that after M iterations, the Markov chain has reached its target distribution. The samples drawn after the burn-in are used for posterior inference.

Thinning, i.e. saving only every kth iteration, is a strategy commonly adopted

in MCMC to reduce high sample autocorrelation and avoid biased Monte Carlo standard errors when consecutive draws are highly correlated with each other. For example, to obtain a run of 10,000 iterations one would run $k \times 10,000$ simulations and save only every kth one. But note that thinning a Markov chain can be unnecessary and inefficient because a fraction of all the posterior samples generated are thrown away [83]. Maceachern and Berliner [87] show that you always get more precise posterior estimates if the entire Markov chain is used. However, thinning is also an attempt to speed up post-processing or reduce required computer storage.

Assessing the convergence of a Markov chain is always a major challenge. Despite much theoretical research into convergence issues, there is limited benefit thus far for practical applications. Although it remains impossible to be completely certain that the simulated draws are representative enough to summarize the posterior distribution or calculate any relevant quantities of interest, there are many diagnostic measures and techniques available to help evaluating convergence. Cowles and Carlin [29] provide a detailed review of convergence diagnostics.

A trace plot of samples versus the simulation index is one of the most straightforward visual analysis in assessing convergence. Two examples of trace plots are shown in Figure 1.4.

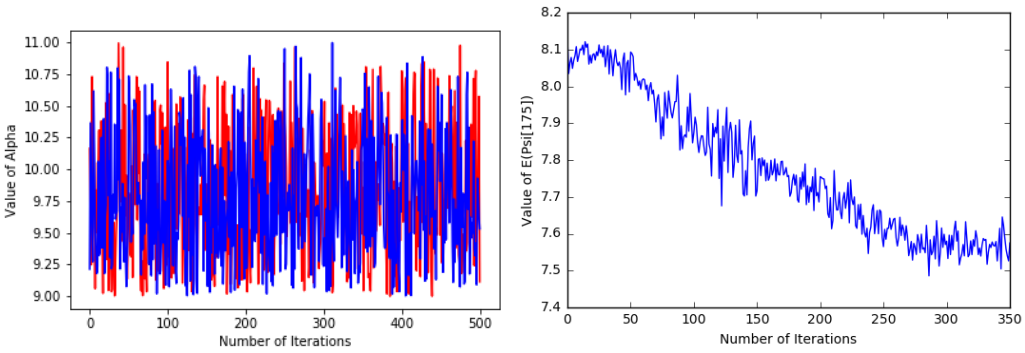

Figure 1.4: Left panel: a trace plot of a well-mixed Markov chain; Right panel: a trace plot of a poorly-mixed Markov chain

The plots in Figure 1.4 indicate whether the chain has converged to its stationary distribution or if it gets stuck in a local region and moves poorly. A relatively constant mean and variance are usually aspects of stationarity that are examined in a trace plot. Plotting the sample mean up to each iteration versus the simulation index can also be useful to detect a relatively constant mean. In addition, investigating the auto-correlation is helpful to assess the convergence of the Markov chain. The decrease of the correlation as the lag (kth lag auto-correlation) increases is an indication of convergence.

There are several Markov chain convergence diagnostic tests, e.g. Gelman-Rubin convergence diagnostic and Geweke diagnostic. The Gelman-Rubin diagnostics evaluates MCMC convergence by analysing the difference between multiple Markov chains. For each model parameter, between-chains and within-chain variances are estimated. Large differences between these variances indicate non-convergence. The Geweke diagnostic compares the means of the first and last part of a Markov chain. If the difference between these two means are small enough, the Markov chain is assumed to have reached the stationary distribution.

Several software packages, e.g. The BUGS [86] (Bayesian inference Using Gibbs Sampling) family and JAGS (Just another Gibbs sampler), are available for automating the Bayesian analysis of models by MCMC methods. JAGS is a program developed by Plummer [102] and has been employed in many fields including metabolomics. JAGS, being compatible with BUGS family by using a particular version of the same modelling language, is highly extensible and allows users to develop new libraries and add-ons.

After it was released in 2012, Stan [21], another software written in C++ for statistical inference, is gaining increasing popularity. Stan uses reverse-mode automatic

differentiation to calculate gradients of the model so that gradient-based MCMC algorithms, gradient-based variational Bayesian methods and gradient-based optimization can be implemented. No-U-Turn sampler (NUTS), Hamiltonian Monte Carlo (HMC), Black-box Variation Inference (BBVI) are some examples of algorithms behind Stan. These algorithms make Stan usually more efficient and converge faster than JAGS.

1.5 Variational Inference

Variational Inference (VI) [72, 133] is a powerful method to approximate intractable integrals. As an alternative strategy to Markov chain Monte Carlo (MCMC) sampling, VI is fast, relatively straightforward for monitoring convergence and typically easier to scale to large data [14] than MCMC. The key idea of VI is to approximate difficult-to-compute conditional densities of latent variables, given observations, via use of optimization. A family of distributions is assumed for the latent variables, as an approximation to the exact conditional distribution. VI aims at finding the member, amongst the selected family, that minimizes the Kullback-Leibler (KL) divergence from the conditional law of interest.

Let x and z denote, respectively, the observed data and latent variables. The goal of the inference problem is to identify the conditional density (assuming a relevant reference measure, e.g. Lebesgue) of latent variables given observations, i.e. $p(z|x)$. Let $\mathscr{L}$ denote a family of densities defined over the space of latent variables – we denote members of this family as $q = q(z)$ below. The goal of VI is to find the element of the family closest in KL divergence to the true $p(z|x)$. Thus, the original inference problem can be rewritten as an optimization one: identify q^* such that

$$q^* = \underset{q \in \mathscr{L}}{\operatorname{argmin}} \, \mathrm{KL}(q \mid p(\cdot|x)) \tag{1.11}$$

for the KL-divergence defined as

$$\begin{aligned}\mathrm{KL}(q \mid p(\cdot|x)) &= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z|x)] \\ &= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z,x)] + \log p(x),\end{aligned}$$

with $\log p(x)$ being constant w.r.t. z. Notation $\mathbb{E}_q$ refers to expectation taken over $z \sim q$. Thus, minimizing the KL divergence is equivalent to maximising the evidence lower bound, $\mathrm{ELBO}(q)$, given by

$$\mathrm{ELBO}(q) = \mathbb{E}_q[\log p(z,x)] - \mathbb{E}_q[\log q(z)]. \tag{1.12}$$

Let $\mathsf{S}_p \subseteq \mathbb{R}^m$, $m \geq 1$, denote the support of the target $p(z|x)$, and $\mathsf{S}_q \subseteq \mathbb{R}^m$ the support of a variational density $q \in \mathscr{L}$ – assumed to be common over all members $q \in \mathscr{L}$. Necessarily, $\mathsf{S}_p \subseteq \mathsf{S}_q$, otherwise the KL-divergence will diverge to $+\infty$.

Many VI algorithms focus on the mean-field variational family, where variational densities in $\mathscr{L}$ are assumed to factorise over blocks of z. That is,

$$q(z) = \prod_{i=1}^{b} q_i(z_i), \quad \mathsf{S}_q = \mathsf{S}_{q_1} \times \cdots \times \mathsf{S}_{q_b}, \quad z = (z_1, \ldots, z_b) \in \mathsf{S}_q,\ z_i \in \mathsf{S}_{q_i}, \tag{1.13}$$

for individual supports $\mathsf{S}_{q_i} \subseteq \mathbb{R}^{m_i}$, $m_i \geq 1$, $1 \leq i \leq b$, for some $b \geq 1$, and $\sum_i m_i = m$. It is advisable that highly correlated latent variables are placed in the same block to improve the performance of the VI method.

There are, in general, two types of approaches to maximise ELBO in VI: a co-ordinate ascent approach and a gradient-based one. Co-ordinate ascent VI (CAVI) [13] is amongst the most commonly used algorithms in this context. To obtain a local maximiser for ELBO, CAVI sequentially optimizes each factor of the mean-field variational density, while holding the others fixed. Analytical calculations on function space – involving variational derivatives – imply that, for given fixed

$q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_b$, ELBO$(q)$ is maximised for

$$q_i(z_i) \propto \exp\left\{\mathbb{E}_{-i}[\log p(z_{i_-}, z_i, z_{i_+}, x)]\right\}, \tag{1.14}$$

where $z_{-i} := (z_{i_-}, z_{i_+})$ denotes vector z having removed component z_i, with i_- (resp. i_+) denoting the ordered indices that are smaller (resp. larger) than i; $\mathbb{E}_{-i}$ is the expectation taken under z_{-i} following its variational distribution, denoted q_{-i}. The above suggest immediately an iterative algorithm, guaranteed to provide values for ELBO(q) that cannot decrease as the updates are carried out.

The expected value $\mathbb{E}_{-i}[\log p(z_{i_-}, z_i, z_{i_+}, x)]$ can be difficult to derive analytically. Also, CAVI typically requires traversing the entire dataset at each iteration, which can be overly computationally expensive for large datasets. Gradient-based approaches, which can potentially scale up to large data – alluding here to recent Stochastic-Gradient-type methods – can be an effective alternative for ELBO optimisation. However, such algorithms have their own challenges, e.g. in the case reparameterization Variational Bayes (VB) analytical derivation of gradients of the log-likelihood can often be problematic, while in the case of score-function VB the requirement of the gradient of $\log q$ restricts the range of the family $\mathscr{L}$ we can choose from.

In real-world applications, hybrid methods combining Monte Carlo with recursive algorithms are common, e.g., Auto-Encoding Variational Bayes, Doubly-Stochastic Variational Bayes for non-conjugate inference, Stochastic Expectation-Maximization (EM) [7, 116, 134]. In VI, Monte Carlo is often used to estimate the expectation within CAVI or the gradient within derivative-driven methods. This is the case, e.g., for Stochastic VI [63] and Black-Box VI (BBVI) [105].

BBVI is used in this work as a representative of gradient-based VI algorithms. It allows carrying out VI over a wide range of complex models. The variational density q is typically chosen within a parametric family, so finding q^* in (1.11) is equiv-

alent to determining an optimal set of parameters that characterize $q_i = q_i(\cdot|\lambda_i)$, $\lambda_i \in \Lambda_i \subseteq \mathbb{R}^{d_i}$, $1 \leq d_i$, $1 \leq i \leq b$, with $\sum_{i=1}^{b} d_i = d$. The gradient of ELBO w.r.t. the variational parameters $\lambda = (\lambda_1, \ldots, \lambda_b)$ equals

$$\nabla_\lambda \text{ELBO}(q) := \mathbb{E}_q\big[\nabla_\lambda \log q(z|\lambda)\{\log p(z,x) - \log q(z|\lambda)\}\big] \tag{1.15}$$

and can be approximated by black-box Monte Carlo estimators as, e.g.,

$$\nabla_\lambda \widehat{\text{ELBO}}(q) := \tfrac{1}{N} \sum_{n=1}^{N} \big[\nabla_\lambda \log q(z^{(n)}|\lambda)\{\log p(z^{(n)},x) - \log q(z^{(n)}|\lambda)\}\big], \tag{1.16}$$

with $z^{(n)} \overset{iid}{\sim} q(z|\lambda)$, $1 \leq n \leq N$, $N \geq 1$. The approximated gradient of ELBO can then be used within a stochastic optimization procedure to update λ at the kth iteration with

$$\lambda_{k+1} \leftarrow \lambda_k + \rho_k \nabla_{\lambda_k} \widehat{\text{ELBO}}(q), \tag{1.17}$$

where $\{\rho_k\}_{k\geq 0}$ is a Robbins-Monro-type step-size sequence [107]. As we will see in later sections, BBVI is accompanied by generic variance reduction methods, as the variability of (1.16) for complex models can be large.

Remark 1 (Hard Constraints)**.** *Though gradient-based VI methods are some times more straightforward to apply than co-ordinate ascent ones, – e.g. combined with the use of modern approaches for automatic differentiation [77] – co-ordinate ascent methods can still be important for models with* hard constraints*, where gradient-based algorithms are laborious to apply. (We adopt the viewpoint here that one chooses variational densities that respect the constraints of the target, for improved accuracy.) Indeed, notice in the brief description we have given above for CAVI and BBVI, the two methodologies are structurally different, as CAVI does not necessarily require to be build up via the introduction of an exogenous variational parameter λ. Thus, in the context of a support for the target $p(z|x)$ that involves*

complex constraints, a CAVI approach overcomes this issue naturally by blocking together the z_i's responsible for the constraints. In contrast, introduction of the variational parameter λ creates sometimes severe complications in the development of the derivative-driven algorithm, as normalising constants that depend on λ are extremely difficult to calculate analytically and obtain their derivatives. Thus, a main argument spanning this work – and illustrated within it – is that co-ordinate-ascent-based VI methods have a critical role to play amongst VI approaches for important classes of statistical models.

Remark 2. *The discussion in Remark 1 is also relevant when VB is applied with constraints imposed on the* variational parameters*. E.g. the latter can involve covariance matrices, whence optimisation has to be carried out on the space of symmetric positive definite matrices. Recent attempts in the VB field to overcome this issue involves updates carried out on manifolds, see e.g. Tran et al. [125].*

Remark 3. *Inserting Monte Carlo steps within a VI approach (that might use a mean field or another approximation) is not uncommon in the VI literature. E.g., Forbes and Fort [48] employ an MCMC procedure in the context of a Variational EM (VEM), to obtain estimates of the normalizing constant for Markov Random Fields – they provide asymptotic results for the correctness of the complete algorithm; Tran et al. [126] apply Mean-Field Variational Bayes (VB) for Generalised Linear Mixed Models, and use Monte Carlo for the approximation of analytically intractable required expectations under the variational densities; several references for related works are given in the above papers. Our work focuses on MC-CAVI, and develops theory that is appropriate for this VI method. This algorithm has* not *been studied analytically in the literature, thus the development of its theoretical justification – even if it borrows elements from Monte Carlo EM – is new.*

1.6 A Bayesian Model of NMR Spectra

The NMR spectrum can contain information for a few hundreds of compounds. Resonance peaks generated by each compound must be identified in the spectrum after deconvolution. The spectral signature of a compound is given by a combination of peaks not necessarily close to each other. Such compounds can generate hundreds of resonance peaks, many of which overlap. This causes difficulty in peak identification and deconvolution. The analysis of NMR spectrum is further complicated by fluctuations in peak positions among spectra induced by uncontrollable variations in experimental conditions and the chemical properties of the biological samples, e.g. by the pH. Nevertheless, extensive information on the patterns of spectral resonance generated by human metabolites is now available in online databases. By incorporating this information into a Bayesian model, we can deconvolve resonance peaks from a spectrum and obtain explicit concentration estimates for the corresponding metabolites. Spectral resonances that cannot be deconvolved in this way may also be of scientific interest; these are modelled in Astle et al. [3] using wavelet basis functions. More specifically, an NMR spectrum is a collection of peaks convoluted with various horizontal translations and vertical scalings, with each peak having the form of a Lorentzian curve. A number of metabolites of interest have known NMR spectrum shape, with the height of the peaks or their width in a particular experiment providing information about the abundance of each metabolite.

We now describe the Bayesian model of NMR Spectra for the deconvolution and quantification of metabolites developed by Astle et al. [3] which is used to compare the efficacy of MCMC and VI in Chapter 3 and inspires our model for 2D JRES spectrum in Chapter 4s. The available data are represented by the pair $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}$ is a vector of n ordered points (of the order $10^3 - 10^4$) on the chemical shift axis – often regularly spaced – and $\mathbf{y}$ is the vector of the corresponding resonance

intensity measurements (scaled, so that they sum up to 1). The conditional law of $\mathbf{y}|\mathbf{x}$ is modelled under the assumption that $y_i|\mathbf{x}$ are independent Normal variables and

$$\mathbb{E}[y_i|\mathbf{x}] = \phi(x_i) + \xi(x_i), \quad 1 \leq i \leq n. \tag{1.18}$$

Here, the ϕ component of the model represents signatures that we wish to assign to target metabolites. The ξ component models signatures of remaining metabolites present in the spectrum, but not explicitly modelled. We refer to this latter as residual spectrum and we highlight the fact that it is important to account for it as it can unveil important information not captured by $\phi(\cdot)$. Function ϕ is constructed parametrically using results from the physical theory of NMR and information available from online databases or expert knowledge, while ξ is modelled semiparametrically with wavelets generated by a mother wavelet (symlet 6) that resembles the Lorentzian curve.

More analytically,

$$\phi(x_i) = \sum_{m=1}^{M} t_m(x_i)\beta_m$$

where M is the number of metabolites modelled explicitly and $\beta = (\beta_1, \ldots, \beta_M)^\top$ is a parameter vector corresponding to metabolite concentrations. Based on the theoretical shape function for NMR peaks (Eq. 1.1), function $t_m(\cdot)$ represents a continuous template function that specifies the NMR signature of metabolite m and it is defined as

$$t_m(\delta) = \sum_{u} \sum_{v=1}^{V_{m,u}} z_{m,u}\, \omega_{m,u,v}\, \ell_\gamma(\delta - \delta^*_{m,u} - c_{m,u,v}), \quad \delta > 0, \tag{1.19}$$

where u is an index running over all multiplets assigned to metabolite m, v is an index representing a peak in a multiplet and $V_{m,u}$ is the number of peaks in multiplet u of metabolite m. In addition, $\delta^*_{m,u}$ specifies the theoretical position on the chemical shift axis of the centre of mass of the uth multiplet of the mth metabolite;

$z_{m,u}$ is a positive quantity, usually equal to the number of protons in a molecule of metabolite m that contributes to the resonance signal of multiplet u; $\omega_{m,u,v}$ is the weight determining the relative heights of the peaks of the multiplet; $c_{m,u,v}$ is the translation determining the horizontal offsets of the peaks from the centre of mass of the multiplet. Both $\omega_{m,u,v}$ and $c_{m,u,v}$ can be computed by empirical estimates of the so-called J-coupling constants; see Hore [67] for more details. The information of $z_{m,u}$ and J-coupling constants can be found in online databases or from expert knowledge.

The residual spectrum is modelled through wavelets,

$$\xi(x_i) = \sum_{j,k} \varphi_{j,k}(x_i)\vartheta_{j,k}$$

where $\varphi_{j,k}(\cdot)$ denote the orthogonal wavelet functions generated by the symlet-6 mother wavelet, see Astle et al. [3] for full details; here, $\vartheta = (\vartheta_{1,1}, \ldots, \vartheta_{j,k}, \ldots)^\top$ is the vector of wavelet coefficients. Indices j,k correspond to the kth wavelet in the jth scaling level.

Finally, overall, the model for an NMR spectrum can be re-written in matrix form as:

$$\mathscr{W}(\mathbf{y} - \mathbf{T}\beta) = \mathbf{I}_{n_1}\vartheta + \varepsilon, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(0, \mathbf{I}_{n_1}/\lambda), \tag{1.20}$$

where $\mathscr{W} \in \mathbb{R}^{n \times n_1}$ is the inverse wavelet transform, M is the total number of known metabolites, $\mathbf{T}$ is an $n \times M$ matrix with its (i,m)th entry equal to $t_m(x_i)$ and λ is a scalar precision parameter.

1.6.1 Prior Specification

Astle et al. [3] assign the following prior distribution to the parameters in the Bayesian model. For the concentration parameters β_m, because its support should be confined to $\mathbb{R}^+$, Astle et al. [3] assume a Normal prior truncated below at zero

for each β_m,

$$\beta_m \sim \text{TN}(e_m, 1/s_m, 0, \infty),$$

where $e_m = 0$ and $s_m = 10^{-3}$, for all $m = 1, \ldots, M$. This prior distribution is flexible enough for a wide range of research problems. Since Astle et al. [3] focus on the spectra generated by biofluids, for which peak widths vary negligibly within spectra, a single common peak width parameter γ is assumed for peaks within a spectrum and the prior distribution is

$$\gamma \sim \text{LN}(0, 1),$$

where LN denotes a Log-Normal distribution. Moreover, for the multiplet chemical shift parameters $\delta^*_{m,u}$, which do fluctuate slightly between spectra due to different experimental conditions, an estimate $\hat{\delta}^*_{m,u}$ for each $\delta^*_{m,u}$ is obtained from online databases (e.g. HMDB [see 136, 137, 138, 139]) to construct an informative prior that

$$\delta^*_{m,u} \sim \text{TN}(\hat{\delta}^*_{m,u}, 10^{-4}, \hat{\delta}^*_{m,u} - 0.03, \hat{\delta}^*_{m,u} + 0.03),$$

where the truncation is chosen because the positional noise is local and smaller fluctuations are more probable. In the regions of the spectrum where both parametric (i.e. ϕ) and semiparametric (i.e. ξ) components need to be fitted, the likelihood is unidentifiable. To tackle this problem, Astle et al. [3] opt for shrinkage priors for the wavelet coefficients and include a vector of hyperparameters ψ – each component $\psi_{j,k}$ of which corresponds to a wavelet coefficient – to penalize the semiparametric component. To reflect prior knowledge that NMR spectra are usually restricted to the half plane above the chemical shift axis, Astle et al. [3] introduce a vector of hyperparameters τ, each component of which, τ_i, corresponds to a spectral data point, to further penalize spectral reconstructions in which some components of

$\mathcal{W}^{-1}\boldsymbol{\vartheta}$ are less than a small negative threshold. In conclusion, motivated by a scale mixture of Multivariate Normals with smoothed truncation limits

$$p(\vartheta|\psi,\tau,\lambda) = \frac{\lambda^{n_1/2}\prod_{j,k}\psi_{j,k}^{1/2}}{C_{\psi,\tau,\lambda}}\exp\left(-\frac{1}{2}\sum_{j,k}\lambda\,\psi_{j,k}\vartheta_{j,k}^2\right)\mathbb{1}\left\{\mathcal{W}^{-1}\vartheta \geq \tau\right\},$$

$$\psi_{j,k} \sim \text{Gamma}(c_j, d_j/2),$$

$$\tau_i \sim \text{TN}(h, 1/(\lambda r), -\infty, h),$$

$$\lambda \sim \text{Gamma}(a, e/2)$$

Astle et al. [3] specify the following joint prior density for $(\vartheta,\psi,\tau,\lambda)$,

$$p(\vartheta,\psi,\tau,\lambda) \propto \lambda^{a+\frac{n+n_1}{2}-1}\Big\{\prod_{j,k}\psi_{j,k}^{c_j-0.5}\exp\big(-\tfrac{\psi_{j,k}d_j}{2}\big)\Big\}$$
$$\times\exp\Big\{-\tfrac{\lambda}{2}\Big(e+\sum_{j,k}\psi_{j,k}\,\vartheta_{j,k}^2+r\sum_{i=1}^{n}(\tau_i-h)^2\Big)\Big\}$$
$$\times\,\mathbb{1}\left\{\mathcal{W}^{-1}\vartheta \geq \tau,\; h\mathbf{1}_n \geq \tau\right\},$$

where ψ introduces local shrinkage for the marginal prior of ϑ and τ is a vector of n truncation limits, which bounds $\mathcal{W}^{-1}\vartheta$ from below. The truncation imposes an identifiability constraint: without it, when the signature template does not match the shape of the spectral data, the mismatch will be compensated by negative wavelet coefficients, such that an ideal overall model fit is achieved even though the signature template is erroneously assigned and the concentration of metabolites is overestimated. Finally we set $c_j = 0.05$, $d_j = 10^{-8}$, $h = -0.002$, $r = 10^5$, $a = 10^{-9}$, $e = 10^{-6}$; see Astle et al. [3] for more details.

1.6.2 MCMC Algorithm

To make inferences about the model parameters, Astle et al. [3] implement an MCMC algorithm with three types of MCMC updates:

- There are Gibbs samplers for β, ϑ, ψ, τ and λ

- There are Metropolis-Hastings updates for each $\delta^*_{m,u}$ and γ
- There are Metropolis-Hastings block updates to break the posterior correlation between the semiparametric (ξ) and the parametric (ϕ) model components. The block updates include: (i) $\delta^*_{m,u}$ and ϑ; (ii) β and ϑ

In addition, in order to improve convergence and mixing, Astle et al. [3] temper the likelihood and penalize the wavelet component of the model during the burn-in stage.

Python, C++ and R have been extensively used during the development of this thesis. In Chapter 2, JAGS was used via R to perform Bayesian statistical analyses. Python was adopted to analyse data from online databases and summarise posterior distributions. In Chapter 3, Python was employed to perform Bayesian inference for relatively simple models and C++ was used for full NMR spectra analysis. In Chapter 4, 1D NMR data and JRES data were analysed with C++.

1.7 Aims

Metabolite identification, data processing and interpretation of results are three major bottlenecks within metabolomic research. Metabolite identification, which is complicated by the great variability of molecular structures and abundance, depends upon the robustness of the data capture techniques. Data processing and reduction techniques are complicated and depend on each laboratory's focus area since different results can be produced from a same dataset through different software and statistical methods by different research groups. All these affect the reproducibility and validation of metabolite analysis. Given the complex nature of the numerous statistical challenges within metabolomics research, we aim to tackle a few of these problems.

- In Chapter 2, to aid the development of better algorithms for ^{1}H NMR data analysis, we use observed NMR chemical shift titration data to estimate the

number of protonation sites, a key parameter in the theoretical relationship between pH and chemical shift. A Bayesian model is developed and fit to the data incorporating theoretical knowledge on chemical shifts, using model selection procedures in a MCMC algorithm.

- In Chapter 3, to improve the computational efficiency of NMR data analysis, we discuss, and then apply a Monte Carlo Co-ordinate Ascent VI (MC-CAVI) algorithm in a sequence of problems of increasing complexity, and study its performance. We also contrast MC-CAVI with MCMC and a representative of derivative-based VI methods – Black Box VI (BBVI) through simulated and real examples, some of which involve hard constraints, which is the major difficulty in the R package "BATMAN" [3]. In the end, we demonstrate MC-CAVI's effectiveness in NMR spectroscopy analysis.

- In Chapter 4, to aid metabolite identification by reducing resonance overlapping, based on a combination of theoretical templates and B-spline tight wavelet frames, we describe a novel Bayesian method for the analysis of JRES datasets from complex biological mixtures. Posterior inference is performed through specially devised Markov chain Monte Carlo methods. We demonstrate the effectiveness of our approach via analyses of datasets from serum and urine.

- In Chapter 5, we discuss the main results and contributions of this project and outline some future research directions.

Chapter 2

Bayesian Estimation of the Number of Protonation Sites from NMR Spectroscopic Data

This chapter has been published in *Metabolomics*. 2018; 14(5): 56. [140].

2.1 Background

In ^{1}H NMR, the chemical shift and multiplicity pattern are characteristics of the metabolite's chemical structure, but are complicated by small sample-to-sample changes in the position of individual resonances due to changes in pH, ionic strength or other physical parameters[45]. While these can be ameliorated to some degree by careful analytical procedures, such as addition of buffers and control of physical conditions, changes in chemical shifts are still present in most NMR metabolomic data sets. Computational approaches to correct these changes, such as alignment, can introduce artefacts and are not able to correct shift changes which swap the ordering of resonances [132]. Chemical shift changes can become a major problem in the statistical analysis of NMR metabolomics data, as they disrupt the linear relationship between NMR intensity at a given position and metabolite abundance [42]. Thus, it becomes important to characterise and model chemical shift changes

(see e.g. Takis et al. [121]), in part to aid construction of better algorithms for data analysis, such as alignment or peak-fitting. Tredwell et al. [127] in 2016 reported titration model parameters such as acid/base limits and pKas for 33 identified metabolites in human urine, as well as titration curves for a further 65 unidentified peaks. A key problem in modelling NMR spectra from untargeted metabolomics is the unknown structure of the molecules giving rise to each resonance, and thus the lack of knowledge of important parameters. In particular, the number of proton binding sites strongly influences the relationship of chemical shift with pH, but has traditionally been hard to infer from titration data alone. To solve this problem, we aim to develop a Bayesian approach to estimate the number of proton binding sites in ^{1}H NMR metabolomics data, without expert knowledge of the molecule's chemical structure.

2.2 Methods

2.2.1 The model

As protonation is usually rapid and reversible on the NMR timescale, the theoretical chemical shift ($\tilde{\delta}$) is a weighted average of the limiting chemical shifts of the unprotonated (δ_A) and the protonated (δ_{HA}) states of the molecule [58, 120].

H. et al. [58] model the theoretical chemical shift as a function of pH and pKa as follows

$$\tilde{\delta} = \frac{\delta_A + \delta_{HA}(10^{(pK_a - pH)})}{1 + 10^{(pK_a - pH)}} \tag{2.1}$$

Szakacs et al. [120] extend this approach to molecules with $q > 1$ protonation sites:

$$\tilde{\delta} = \frac{\delta_A + \sum_{i=1}^{q} \delta_{HiA} 10^{(\sum_{j=q-i+1}^{q} pK_j) - ipH}}{1 + \sum_{k=1}^{q} 10^{(\sum_{l=q-k+1}^{q} pK_l) - kpH}}, \tag{2.2}$$

accounting for the interaction between protons bound at different binding sites and the statistics of proton binding.

From (2.1)-(2.2), it is evident that the theoretical chemical shift follows a titration curve which describes the position of the resonance over a range of pH. When the number of sites is known, nonlinear fitting can be applied using Equation (2.2) to model the titration curve to obtain the pKa values, as well as the acid and base chemical shift limits [127]. However, in many metabolomics applications (for example alignment), the number of protonation sites may not be known, especially for unknowns or molecules of complex structure. Thus it is of interest to consider whether the number of protonation sites can be estimated along with the pH dependence of the chemical shift.

Here, we focus on inferring the number of protonation sites from observations of chemical shift changes for a given resonance at different pH values. Due to their small size, few metabolites have many protonation sites. We therefore limit the search space to 1-site, 2-site and 3-site models, although the approach can be easily extended to include more than 3 protonation sites if required. We employ a Bayesian approach because it provides a natural way of incorporating prior information and combining results of different experiments. In the Bayesian framework, it is, in principle, easy to incorporate model choice in the inferential process by specifying an appropriate prior distribution on the model space. Posterior inference is performed through Markov chain Monte Carlo (MCMC) methods. In this context, as model selection involves models with different dimensions, we employ a Reversible jump MCMC algorithm, which is implemented in the software JAGS [102].

We propose a non-linear Bayesian regression model for each NMR resonance for

each molecule. In particular, we assume that the observed chemical shift, y_i, follows a Normal distribution, with mean $\tilde{\delta}$, representing the theoretical chemical shift, and variance σ^2, the measurement error:

$$y_i|\tilde{\delta}_i, \sigma^2 \sim N(\tilde{\delta}, \sigma^2)$$

The theoretical chemical shift $\tilde{\delta}_i$ is a function of the pH, pKa, δ_A and the number of protonation sites as described in Equation (2.2).

2.2.2 Specification of Prior Knowledge

Since most metabolites have up to three protonation sites, we specify as prior distribution on the number of protonation sites a Uniform distribution on the set $\{1,2,3\}$. Therefore, each model is a priori equally likely. We complete the model by specifying a prior distribution on the remaining parameters. Assuming no additional spectral effects and conditioning on the number of sites q, we choose a Uniform distribution defined over the NMR ppm scale $[0,10]$ as prior for δ_A and $\delta_{H_jA}, j = 1,\ldots,q$.

Moreover, to improve efficiency in searching the parameter space and avoid identifiability issues (where different combinations of parameter values lead to the same likelihood value so that the model is not able to distinguish between them) we impose an order constraint on the $\delta_{\mathbf{A}}$ and $\delta_{\mathbf{H_jA}}$ values, in descending or ascending order according to the trend of the data. This improves MCMC convergence and the accuracy of estimation. The order direction can be estimated, for example, by fitting a simple linear regression, $\mathbf{y} = \beta\mathbf{pH} + b$, to the data and considering the sign of the estimated slope parameter β. If $\beta > 0$, the relationship between chemical shift and pH is approximately increasing and we would impose the constraint $\delta_A > \delta_{H_1A} > \delta_{H_2A} > \ldots > \delta_{H_qA}$ on the parameter space. On the other hand, if $\beta < 0$,

we would impose restriction $\delta_A < \delta_{H_1A} < \delta_{H_2A} < \ldots < \delta_{H_qA}$.

For most metabolites the change in chemical shift between adjacent protonation sites is smaller than 1ppm and the total shift change from most acidic to most basic peak position is also smaller than 1ppm. This allows us to assume that the change of chemical shift between adjacent protonation sites is smaller than 1ppm, i.e.

$$0 < |\delta_A - \delta_{H_1A}| < 1 \quad 0 < |\delta_{H_jA} - \delta_{H_{j+1}A}| < 1, \ \ j = 1,2,3,...$$

Finally, an Inverse-Gamma prior distribution with parameters (a^2, a), which is often used as a Bayesian prior for error variance, is chosen for σ^2. Note that a, which reflects the measurement error, should be chosen carefully according to the experiment. In our model, $a = 10^4$ is chosen based on empirical estimation of the measurement error related to the resolution of the spectrometer and its ability to measure peak position [74].

The details regarding model parameters are shown in Table 2.1.

Observed	To be estimated
y, pH	σ, $\tilde{\delta}$, q, δ_A, δ_{H_jA}, pKa

Table 2.1: Details about Parameters

We fit the model to each resonance independently. We pick as an estimate of q the number of protonation sites with highest posterior probability. We then refit the same model but fix q equal to its posterior estimate to obtain an estimate of the other parameters conditional on q. Posterior inference is performed in JAGS, running four chains of the MCMC algorithm for 50,000 iterations with a burn-in period of 25,000. The code of JAGS can be found in Appendix A1.

2.2.3 Prior Specification for pKa

A great advantage of working in a Bayesian framework is the ability of the model to incorporate problem specific prior information. To specify informative prior knowledge on the pKa range, which aids computational stability and improves convergence of the MCMC algorithm, we exploit information available in the Human Metabolome Database (version 4.0) [136, 137, 138, 139], which records the pKa values of many common metabolites.

pKa is the negative base-10 logarithm of the acid dissociation constant of a solution. In our modelling, it is the value of pH corresponds to the middle point of the change part of a chemical shift titration curve. By studying the empirical distribution of the pKa values downloaded from the database, we found that the distribution of pKa values has a heavy right tail. We choose as prior range for pKa $[1.2, 13.7]$ to correspond to the pH range of our data. This range includes most metabolites reported in HMDB, but excludes values below the 7% and above the 90% percentile of the pKa distribution.

2.2.4 Data

Details of sample collection, NMR acquisition and data processing can be found in Tredwell et al. [127]. All data used in this study is publicly available as Supplementary material to the original article under the Creative Commons attribution 4.0 International License https://creativecommons.org/licenses/by/4.0/. Briefly, a urine sample was collected from five different individuals and pooled to obtain an average representative human urine sample. To avoid chemical shift effects from metal ions the urine was treated with chelex resin to reduce both Ca^{2+} and Mg^{2+} concentrations without significantly altering metabolite composition. Note that, while this results in non-physiological concentrations of these ions, it is not expected to affect the ability of the model to recover the number of protonation sites. The pool was

then titrated to produce 51 samples covering the range $2 < \texttt{pH} < 12$. Spectra were acquired on a Bruker Avance DRX600 NMR spectrometer (Bruker BioSpin, Rheinstetten, Germany), with a ^{1}H frequency of 600 MHz. A one-dimensional NOESY sequence was used with water suppression, and 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 0.5 Hz. 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) as an internal standard. Metabolite peak positions were obtained using in-house MATLAB scripts. A version of the scripts for peak picking and spline fits are part of the BATMAN project (`batman.r-forge-project.org`). Data for one metabolite (phenylalanine at 7.35 and 7.41 ppm) were discarded as it was found that the peak positions could not be measured accurately due to the high level of peak overlap in this region of the spectra.

2.3 Results and Discussion

Our aim is to estimate the number of protonation sites for small molecule metabolites from their observed NMR pH titration curves. From Figure 2.1, it is clear that when the number of protonation sites is estimated correctly, the chemical shift changes match the data quite well.

Table 2.2: Comparison of the literature number of sites and the number estimated by the model

		Estimated Number of Sites			Total
		1	2	3	
Literature Number of Sites	1	**25**	1	0	26
	2	5	**9**	0	14
	3	0	4	**7**	11
Total		30	14	7	51

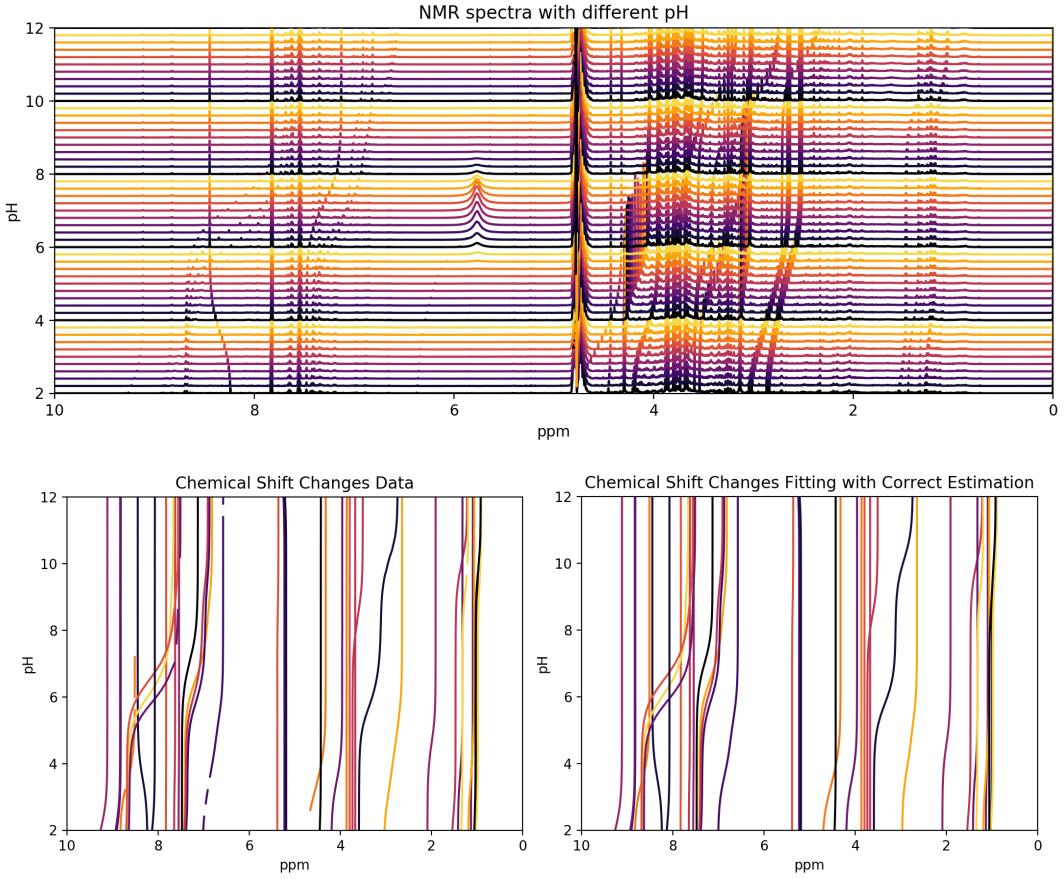

Figure 2.1: Upper panel: ^{1}H NMR spectra with pH adjusted from 2 to 12. Lower left panel: Observed chemical shift positions (y) of 51 resonances. Lower right panel: Fitted chemical shift positions ($\tilde{\delta}$) for the 51 resonances. Only resonances with correct q predicted are shown.

A summary of the results is shown in Table 2.2. More detailed results for each resonance can be found in Table 2.3. Of the 51 resonances, the estimated number of sites matches that found in the literature in 41 cases (80.4%). It is evident that most of the incorrect predictions (10 out of 51) result from an underestimation of the number of sites compared to the literature value. The literature site numbers are sourced from Handbook of Biochemistry and Molecular Biology [85]. Where this was not possible, (Hydroxyisobutyrate, Hydroxyisovalerate, Indoxyl and Methyl-2-Oxovalerate) the number was determined from an assessment of the molecular structure.

Table 2.3: Probability of different numbers of protonation sites, estimated number of protonation sites and literature number of protonation sites for 51 resonances from 32 metabolites in human urine. Rows with correctly estimated numbers of sites are shown in bold.

Metabolite	Database ID	Chemical Shift at PH7.4	1 Site Prob.	2 Site Prob.	3 Site Prob.	Estimated Number of Sites	Literature Number of Sites
Hydroxy-isobutyrate	**HMDB0000729**	**1.347**	**91.893**	**7.488**	**0.619**	**1**	**1**
Hydroxy-isovalerate	**HMDB0000754**	**1.260**	**86.597**	**12.795**	**0.608**	**1**	**1**
Indoxyl	**HMDB0004094**	**7.192**	**93.017**	**5.688**	**1.295**	**1**	**1**
Methyl-2-Oxovalerate	**HMDB0000695**	**1.093**	**95.383**	**4.326**	**0.291**	**1**	**1**
Acetate	**HMDB0000042**	**1.910**	**93.326**	**5.879**	**0.795**	**1**	**1**
Alanine	**HMDB0000161**	**1.212**	**0**	**80.827**	**19.173**	**2**	**2**
Allantoin	**HMDB0000462**	**5.383**	**97.868**	**2.049**	**0.083**	**1**	**1**
Citrate	HMDB0000094	2.528	0	75.404	24.596	2	3
Citrate	**HMDB0000094**	**2.646**	**0**	**47.869**	**52.131**	**3**	**3**
Creatinine	HMDB0000562	3.033	87.889	11.593	0.518	1	2
Creatinine	HMDB0000562	4.043	94.992	4.65	0.358	1	2
Formate	**HMDB0000142**	**8.448**	**92.786**	**6.377**	**0.837**	**1**	**1**
Glucose	**HMDB0000122**	**5.228**	**98.661**	**1.284**	**0.055**	**1**	**1**
Hippurate	**HMDB0000714**	**3.960**	**70.111**	**27.403**	**2.486**	**1**	**1**
Hippurate	**HMDB0000714**	**7.541**	**86.036**	**9.798**	**4.166**	**1**	**1**
Hippurate	**HMDB0000714**	**7.627**	**92.742**	**6.114**	**1.144**	**1**	**1**
Hippurate	**HMDB0000714**	**7.824**	**92.264**	**5.387**	**2.349**	**1**	**1**
Hippurate	**HMDB0000714**	**8.512**	**54.082**	**26.841**	**19.077**	**1**	**1**
Histidine	**HMDB0000177**	**7.253**	**0**	**42.669**	**57.331**	**3**	**3**
Histidine	**HMDB0000177**	**8.188**	**0**	**7.55**	**92.45**	**3**	**3**
Imidazole	**HMDB0001525**	**7.229**	**59.201**	**37.952**	**2.847**	**1**	**1**
Imidazole	HMDB0001525	8.040	0	73.582	26.418	2	1
Isoleucine	**HMDB0000172**	**0.902**	**0.801**	**65.964**	**33.235**	**2**	**2**

Lactate	**HMDB0000190**	**1.320**	**89.155**	**9.941**	**0.904**	**1**	**1**
Leucine	HMDB0000687	0.932	83.263	14.099	2.638	1	2
Mannitol	**HMDB0000765**	**3.673**	**92.487**	**6.501**	**1.012**	**1**	**1**
Mannitol	**HMDB0000765**	**3.797**	**96.633**	**2.676**	**0.691**	**1**	**1**
Mannitol	**HMDB0000765**	**3.864**	**96.567**	**2.964**	**0.469**	**1**	**1**
Methyl-Succinate	**HMDB0001844**	**1.062**	**43.561**	**50.274**	**6.165**	**2**	**2**
Piperazine	**HMDB0014730**	**3.526**	**0**	**65.255**	**34.745**	**2**	**2**
TMethyl-Histidine	**HMDB0000479**	**6.873**	**0**	**14.792**	**85.208**	**3**	**3**
TMethyl-Histidine	**HMDB0000479**	**8.306**	**0**	**0**	**100**	**3**	**3**
TMethyl-Histidine	HMDB0000001	3.788	0	94.773	5.227	2	3
TTMethyl-Histidine	**HMDB0000001**	**6.909**	**0**	**16.988**	**83.012**	**3**	**3**
TTMethyl-Histidine	**HMDB0000001**	**8.396**	**0**	**23.514**	**76.486**	**3**	**3**
Tartrate	**HMDB0029878**	**4.322**	**5.764**	**51.864**	**42.372**	**2**	**2**
Taurine	HMDB0000251	3.412	86.863	7.137	6	1	2
Threonine	**HMDB0000167**	**1.194**	**14.472**	**67.882**	**17.646**	**2**	**2**
Trigonelline	**HMDB0000875**	**4.429**	**68.693**	**19.481**	**11.826**	**1**	**1**
Trigonelline	**HMDB0000875**	**8.073**	**74.875**	**17.025**	**8.1**	**1**	**1**
Trigonelline	**HMDB0000875**	**8.822**	**77.588**	**15.531**	**6.881**	**1**	**1**
Trigonelline	**HMDB0000875**	**8.834**	**58.857**	**30.807**	**10.336**	**1**	**1**
Trigonelline	**HMDB0000875**	**9.115**	**67.411**	**21.353**	**11.236**	**1**	**1**
Tris	**CHEBI:9754**	**3.715**	**94.453**	**5.363**	**0.184**	**1**	**1**
Tryptophan	HMDB0000929	7.719	87.38	11.156	1.464	1	2
Tyrosine	HMDB0000158	6.885	0	90.202	9.798	2	3
Tyrosine	HMDB0000158	7.207	0	90.592	9.408	2	3
Valine	**HMDB0000883**	**0.906**	**0**	**79.851**	**20.149**	**2**	**2**
Valine	**HMDB0000883**	**1.060**	**2.967**	**77.568**	**19.465**	**2**	**2**
Xylose	**HMDB0000098**	**5.190**	**98.475**	**1.476**	**0.049**	**1**	**1**

transAconitate	**HMDB0000958**	**6.574**	**0**	**64.477**	**35.523**	**2**	**2**

Given the estimation of the number of protonation sites, the other parameters of the model (acid limits, base limits and pKa values) can be estimated using the same model. The modelled pKa values closely agree with the literature values [85], and the modelled acid and base limits are also in good agreement with the previously modelled values [127]. Therefore we do not present these in detail here. Four examples including 1, 2 and 3 protonation sites, (Acetate, Alanine, Threonine and TTMethylHistidine) are shown in Table 2.4 and Figure 2.2.

Table 2.4: Literature and Modelled Results of Acetate, Alanine, Threonine and TTMethylHistidine

Metabolite	Literature pKa Values			Modelled pKa Values			Modelled Acid and Base Limits			
Acetate	4.760			4.591			1.910	2.089		
Alanine	2.340	9.690		2.384	9.980		1.212	1.472	1.573	
Threonine	2.630	10.430		2.072	9.195		1.194	1.322	1.379	
TTMethyl-Histidine	1.690	6.480	8.850	1.832	6.062	9.302	6.910	7.040	7.390	7.491

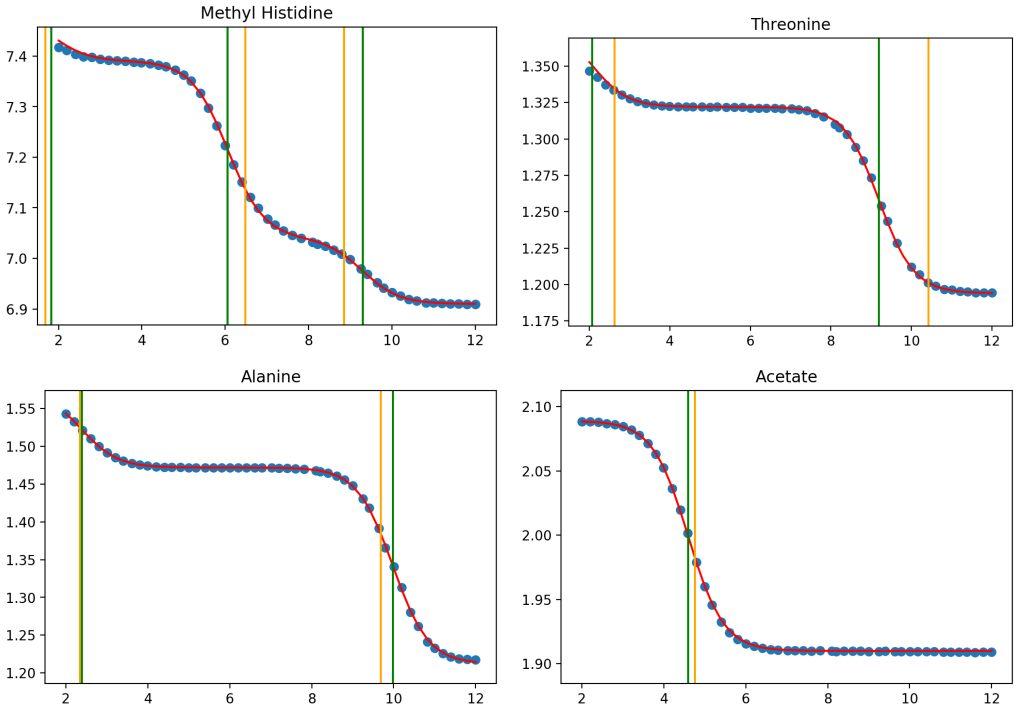

Figure 2.2: Measured Chemical Shift Changes for Acetate, Alanine, Threonine and TTMethylHistidine with literature pKa Values (yellow vertical line), fitted pKa values (green vertical line) and the fit of the theoretical model (red line). The x-axis corresponds to pH and y-axis corresponds to ppm.

2.3.1 Metabolites with incorrectly estimated number of protonation sites

The model failed to estimate the correct number of protonation sites for 10 out of 51 resonances. There are several types of problem leading to incorrect estimation of the number of protonation sites. The first type occurs when at least one literature pKa value lies outside the range of the observed data. Taurine is a good example of this, as shown in Figure 2.3, where it can be seen that one pKa lies at pH 1.5, while the data only cover the pH range 3.2-12.

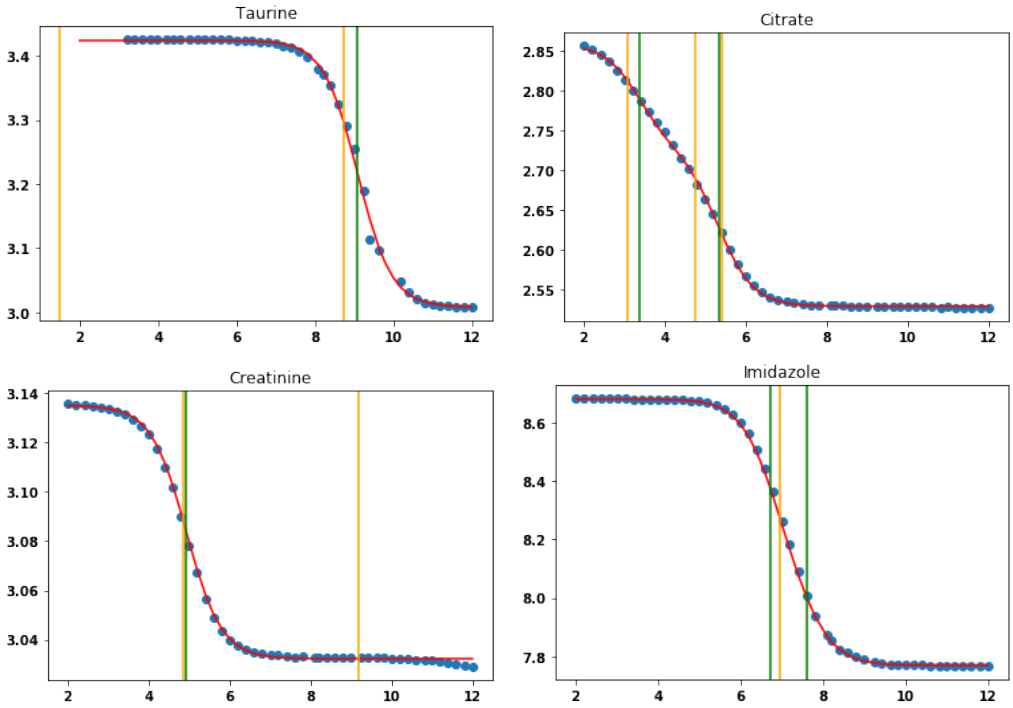

Figure 2.3: Examples of resonances with incorrectly estimated numbers of sites: Taurine, Citrate, Creatinine, Imidazole with literature pKa Values (yellow vertical line), fitted pKa values (green vertical line) and the fit of the theoretical model (red line). The x-axis corresponds to pH and y-axis corresponds to ppm.

The second type of inaccurate estimation happens when two adjacent pKas are so close that the change in chemical shift between them is too small compared to the measurement error. The δ 2.7 resonance of citrate is a good example of this, as in Figure 2.3, where the smooth titration curve around pH 4-5 does not suggest the presence of the third pKa at 4.75. A third type of incorrect estimate happens when the change of chemical shift is too small so that the transition can not be detected near the pKa value, for instance creatinine as shown in Figure 2.3. Conversely, the change in the chemical shift can be too large compared to the estimated measurement error, for example imidazole as shown in Figure 2.3, forming a fourth type of inaccuracy.

Some molecules have multiple resonances and so the question arises of whether

to combine them, or if not, how to pick the best resonance to model. We do not recommend to combine resonances from the same molecule as, with our data, this tended to over estimate the number of protonation sites leading to a poorer fit. Instead, it is preferred to pick a resonance with "good behaviour", i.e. one which is not overlapped, shows strong changes in chemical shift, but with a good number of observations near each chemical shift transition (near the pKa). When more than one resonance from the same molecule are modelled and give different predictions for the number of sites, we recommend to use information such as the model fit error to judge which estimation is more reliable. We note that this does not apply in fully untargeted analysis when the metabolites are unidentified, and thus one does not know if two resonances come from the same molecule.

2.4 Conclusions

The Bayesian fit based on the model of Szakacs et al. [120] can effectively estimate the number of protonation sites for many small molecule metabolites, given sufficient pH titration data. Incorrect estimations are mainly due to cases where pKa values are very similar, and thus could not be distinguished, and/or a lack of data in the necessary pH ranges. We note that, even when the number of sites was incorrectly estimated, it is still possible to estimate the chemical shift position of a resonance quite accurately in most cases. The information obtained from the modelling procedure described here could be useful in a number of ways. For example, the pH could be estimated from the positions of a few well known and easily located resonances. This could then be used to predict the chemical shift positions of resonances of other metabolites expected in a sample, which could then help with automated annotation, alignment or peak fitting (as an initial position estimate). The predicted number of protonation sites may also be helpful during the process of identifying unknown compounds, although orthogonal analytical information would almost al-

ways be needed in addition. Overall, we hope that this modelling approach may be valuable for the future development of algorithms for analysis of metabolomic ^{1}H NMR spectra including alignment, annotation and peak fitting.

2.5 Compliance with Ethical Standards

Ethical approval This study analysed previously collected data which involved human participants who had provided informed consent. These ethical issues are described in detail in Tredwell et al. [127].

Informed Consent Informed consent was obtained from all individual participants included in the study.

Data Availability The metabolomics and metadata reported in this paper are available as supplementary information to the original study [127] which is available from the Springer website under the Creative Commons attribution 4.0 International License https://creativecommons.org/licenses/by/4.0/.

Chapter 3

On the Efficacy of Monte Carlo Implementation of CAVI

3.1 Background

Sophisticated statistical techniques are essential for NMR data analysis because NMR data is often large and heavily structured . Astle et al. [3] incorporate extensive information on the patterns of spectral resonance generated by human metabolites from online databases into a Bayesian model and deconvolve resonance peaks from a spectrum and obtain explicit concentration estimates for the corresponding metabolites. Posterior inference is performed by MCMC with specifically designed block updates and annealing. Variational Inference (VI) [72, 133] is a powerful alternative strategy to MCMC sampling because it is fast, straightforward for monitoring convergence and typically easier to scale to large data [14] than MCMC. In real-world implementations of complex models, Monte Carlo methods are widely used to estimate expectations in coordinate-ascent VI algorithms and gradients in derivative-driven VI algorithms. Therefore, combing Monte Carlo methods with VI has great potential to improve computational efficiency of NMR data analysis.

The main contributions of this section are:

(i) We discuss, and then apply a Monte Carlo CAVI (MC-CAVI) algorithm in a sequence of problems of increasing complexity, and study its performance. As the name suggests, MC-CAVI uses the Monte Carlo principle for the approximation of difficult-to-compute conditional expectations, $\mathbb{E}_{-i}[\log p(z_{i_-}, z_i, z_{i_+}, x)]$, within CAVI.

(ii) We provide a theoretical justification for the algorithm by showing analytically that, under suitable regularity conditions, MC-CAVI will get arbitrarily close to a maximiser of the ELBO with high probability.

(iii) We contrast MC-CAVI with MCMC and BBVI through simulated and real examples, some of which involve hard constraints; we demonstrate MC-CAVI's effectiveness in an important application imposing such hard constraints, with real data in the context of Nuclear Magnetic Resonance (NMR) spectroscopy.

3.2 MC-CAVI Algorithm

3.2.1 Description of the Algorithm

We begin with a description of the basic CAVI algorithm. A double subscript will be used to identify block variational densities: $q_{i,k}(z_i)$ (resp. $q_{-i,k}(z_{-i})$) refers to the density of the ith block (resp. all blocks but the ith), after k updates have been carried out on that block density (resp. k updates have been carried out on the blocks preceding the ith, and $k-1$ updates on the blocks following the ith).

- Step 0: Initialize probability density functions $q_{i,0}(z_i)$, $i = 1, \ldots, b$.
- Step k: For $k \geq 1$, given $q_{i,k-1}(z_i)$, $i = 1, \ldots, b$, execute:
  - For $i = 1, \ldots, b$, update:

$$\log q_{i,k}(z_i) = const. + \mathbb{E}_{-i,k}[\log p(z,x)],$$

with $\mathbb{E}_{-i,k}$ taken w.r.t. $z_{-i} \sim q_{-i,k}$.

- Iterate until convergence.

Assume that the expectations $\mathbb{E}_{-i}[\log p(z,x)]$, $\{i : i \in \mathscr{I}\}$, for an index set $\mathscr{I} \subseteq \{1,\ldots,b\}$, can be obtained analytically, over all updates of the variational density $q(z)$; and that this is not the case for $i \notin \mathscr{I}$. Intractable integrals can be approximated via a Monte Carlo method. (As we will see in the applications later in the chapter, such a Monte Carlo device typically uses samples from an appropriate MCMC algorithm.) In particular, for $i \notin \mathscr{I}$, one obtains $N \geq 1$ samples from the current $q_{-i}(z_{-i})$ and uses the standard Monte Carlo estimate

$$\widehat{\mathbb{E}}_{-i}[\log p(z_{i_-}, z_i, z_{i_+}, x)] = \frac{\sum_{n=1}^{N} \log p(z_{i_-}^{(n)}, z_i, z_{i_+}^{(n)}, x)}{N}.$$

Implementation of such an approach gives rise to MC-CAVI, described in Algorithm 1.

Algorithm 1: MC-CAVI

Require: Number of iterations T and Number of Monte Carlo samples N.

Require: $\mathbb{E}_{-i}[\log p(z_{i_-}, z_i, z_{i_+}, x)]$ in closed form, for $i \in \mathscr{I}$.

Initialize $q_{i,0}(z_i)$, $i = 1,\ldots,b$.

for $k = 1 : T$ **do**

  for $i = 1 : b$ **do**

    If $i \in \mathscr{I}$, set $q_{i,k}(z_i) \propto \exp\left\{\mathbb{E}_{-i,k}[\log p(z_{i_-}, z_i, z_{i_+}, x)]\right\}$;

    If $i \notin \mathscr{I}$:

    Obtain N samples, $(z_{i_-,k}^{(n)}, z_{i_+,k-1}^{(n)})$, $1 \leq n \leq N$, from $q_{-i,k}(z_{-i})$.

    Set

$$q_{i,k}(z_i) \propto \exp\left\{\hat{\mathbb{E}}_{-i,k}[\log p(z_{i_-}, z_i, z_{i_+}, x)]\right\}$$
$$= \exp\left\{\frac{\sum_{n=1}^{N} \log p(z_{i_-,k}^{(n)}, z_i, z_{i_+,k-1}^{(n)}, x)}{N}\right\}.$$

  end

end

Please note that, in Algorithm 1, due to Monte Carlo updates of parameters $\notin \mathscr{I}$, the $\mathbb{E}_{-i,k}[\log p(z_{i_-}, z_i, z_{i_+}, x)]$ is not the exact value of expectation over iterations. We use symbol $\mathbb{E}$ to denote the closed form of expectation while $\hat{\mathbb{E}}$ denotes the expectations without closed forms where we will apply the Monte Carlo techniques. Further illustration is provided in Section 3.3.

3.2.2 Applicability of MC-CAVI

We discuss here the class of problems to which MC-CAVI can be applied. It is desirable to avoid settings where the order of samples or statistics to be stored in memory increases with the iterations of the algorithm. To set-up the ideas we begin with CAVI itself. Motivated by the standard exponential class of distributions, we work as follows.

Consider the case when the target density $p(z,x) \equiv f(z)$ is assumed to have the structure,

$$f(z) = h(z) \exp\big\{\langle \eta, T(z)\rangle - A(\eta)\big\}, \quad z \in \mathsf{S}_p, \tag{3.1}$$

for s-dimensional constant vector $\eta = (\eta_1, \ldots, \eta_s)$, vector function $T(z) = (T_1(z), \ldots, T_s(z))$, with some $s \geq 1$, and relevant scalar functions $h > 0$, A; $\langle \cdot, \cdot \rangle$ is the standard inner product in $\mathbb{R}^s$. Notice that we omit reference to the data x in (3.1), as x is fixed and irrelevant for our purposes and f is not required to integrate to 1. Also, we are given the choice of block-variational densities $q_1(z_1), \ldots, q_b(z_b)$ in (1.13). Following the definition of CAVI from Section 3.2.1 – assuming that the algorithm can be applied, i.e. all required expectations can be obtained analytically – the number of 'sufficient' statistics, say $T_{i,k}$ giving rise to the definition of $q_{i,k}$ will always be upper bounded by s. Thus, in our working scenario, CAVI will be applicable with a computational cost that is upper bounded by a constant within

the class of target distributions in (3.1) – assuming relevant costs for calculating expectations remain bounded over the algorithmic iterations.

Moving on to MC-CAVI, following the definition of index set $\mathscr{I}$ in Section 3.2.1, recall that a Monte Carlo approach is required when updating $q_i(z_i)$ for $i \notin \mathscr{I}$, $1 \leq i \leq b$. In such a scenario, controlling computational costs amounts to having a target (3.1) admitting the factorisations,

$$h(z) \equiv h_i(z_i)h_{-i}(z_{-i}), \quad T_l(z) \equiv T_{l,i}(z_i)T_{l,-i}(z_{-i}),\ 1 \leq l \leq s, \quad \text{for all } i \notin \mathscr{I}. \tag{3.2}$$

Once (3.2) is satisfied, we do not need to store all N samples from $q_{-i}(z_{-i})$, but simply some relevant averages keeping the cost per iteration for the algorithm bounded. We stress that the combination of characterisations in (3.1)-(3.2) is very general and will typically be satisfied for most practical statistical models.

3.2.3 Theoretical Justification of MC-CAVI

An attractive feature of MC-CAVI versus derivative-driven VI methods is its structural similarity with Monte Carlo Expectation-Maximization (MCEM). Thus, one can build on results in the MCEM literature to prove asymptotical properties of MC-CAVI; see e.g. [26, 15, 78, 50]. To avoid technicalities related with working on general spaces of probability density functions, we begin by assuming a parameterised setting for the variational densities – as in the BBVI case – with the family of variational densities being closed under CAVI or (more generally) MC-CAVI updates.

Assumption 1 (Closedness of Parameterised $q(\cdot)$ Under Variational Update)**.** *For the CAVI or the MC-CAVI algorithm, each $q_{i,k}(z_i)$ density obtained during the iterations of the algorithm, $1 \leq i \leq b$, $k \geq 0$, is of the parametric form*

$$q_{i,k}(z_i) = q_i(z_i|\lambda_i^k),$$

for a unique $\lambda_i^k \in \Lambda_i \subseteq \mathbb{R}^{d_i}$*, for some* $d_i \geq 1$*, for all* $1 \leq i \leq b$*.*

(Let $d = \sum_{i=1}^{b} d_i$ *and* $\Lambda = \Lambda_1 \times \cdots \times \Lambda_b$*.)*

Under Assumption 1, CAVI and MC-CAVI can be corresponded to some well-defined maps $M : \Lambda \mapsto \Lambda$, $\mathscr{M}_N : \Lambda \mapsto \Lambda$ respectively, so that, given current variational parameter λ, one step of the algorithms can be expressed in terms of a new parameter λ' (different for each case) obtained via the updates

$$\text{CAVI: } \lambda' = M(\lambda); \qquad \text{MC-CAVI: } \lambda' = \mathscr{M}_N(\lambda).$$

For an analytical study of the convergence properties of CAVI itself and relevant regularity conditions, see e.g. [8, Proposition 2.7.1], or extensive work in numerical optimisation. Expressing the MC-CAVI update – say, the $(k+1)$th one – as

$$\lambda^{k+1} = M(\lambda^k) + \{\mathscr{M}_N(\lambda^k) - M(\lambda^k)\}, \tag{3.3}$$

it can be seen as a random perturbation of a CAVI step. In the rest of this section we will explore the asymptotic properties of MC-CAVI. We follow closely the approach in [26] – as it provides a less technical procedure, compared e.g. to [50] or other work about MCEM – making all appropriate adjustments to fit the derivations into the setting of the MC-CAVI methodology. We denote by M^k, $\mathscr{M}_N^k$, the k-fold composition of M, $\mathscr{M}_N$ respectively, for $k \geq 0$.

Assumption 2. Λ *is an open subset of* $\mathbb{R}^d$*, and the mappings* $\lambda \mapsto ELBO(q(\lambda))$*,* $\lambda \mapsto M(\lambda)$ *are continuous on* Λ*.*

If $M(\lambda) = \lambda$ for some $\lambda \in \Lambda$, then λ is a fixed point of $M()$. A given $\lambda^* \in \Lambda$ is called an isolated local maximiser of the ELBO$(q(\cdot))$ if there is a neighborhood of λ^* over which λ^* is the unique maximiser of the ELBO$(q(\cdot))$.

Assumption 3 (Properties of $M(\cdot)$ Near a Local Maximum)**.** *Let* $\lambda^* \in \Lambda$ *be an isolated local maximum of* $ELBO(q(\cdot))$*. Then,*

(i) λ^* *is a fixed point of* $M(\cdot)$*;*

(ii) there is a neighborhood $V \subseteq \Lambda$ *of* λ^* *over which* λ^* *is a unique maximum, such that* $ELBO(q(M(\lambda))) > ELBO(q(\lambda))$ *for any* $\lambda \in V \backslash \{\lambda^*\}$.

Notice that the above assumption refers to the deterministic update $M(\cdot)$, which performs co-ordinate ascent; thus requirements (i), (ii) are fairly weak for such a recursion. The critical technical assumption required for delivering the convergence results in the rest of this section is the following one.

Assumption 4 (Uniform Convergence in Probability on Compact Sets)**.** *For any compact set* $C \subseteq \Lambda$ *the following holds: for any* $\rho, \rho' > 0$*, there exists a positive integer* N_0*, such that for all* $N \geq N_0$ *we have,*

$$\inf_{\lambda \in C} \text{Prob}\left[\left|\mathscr{M}_N(\lambda) - M(\lambda)\right| < \rho\right] > 1 - \rho'.$$

It is beyond the context of this paper to examine Assumption 4 in more depth. We will only stress that Assumption 4 is the sufficient structural condition that allows to extend closeness between CAVI and MC-CAVI updates in a single algorithmic step into one for arbitrary number of steps.

We continue with a definition.

Definition 1. *A fixed point* λ^* *of* $M(\cdot)$ *is said to be asymptotically stable if,*

(i) for any neighborhood V_1 *of* λ^**, there is a neighborhood* V_2 *of* λ^* *such that for all* $k \geq 0$ *and all* $\lambda \in V_2$*,* $M^k(\lambda) \in V_1$*;*

(ii) there exists a neighbourhood V *of* λ^* *such that* $\lim_{k\to\infty} M^k(\lambda) = \lambda^*$ *if* $\lambda \in V$.

We will state the main asymptotic result for MC-CAVI in Theorem 1 that follows; first we require Lemma 1.

Lemma 1. *Let Assumptions 1-3 hold. If* λ^* *is an isolated local maximiser of* $ELBO(q(\cdot))$*, then* λ^* *is an asymptotically stable fixed point of* $M(\cdot)$.

The main result of this section is as follows.

Theorem 1. *Let Assumptions 1-4 hold and λ^* be an isolated local maximiser of* $\mathrm{ELBO}(q(\cdot))$. *Then there exists a neighbourhood, say V_1, of λ^* such that for starting values $\lambda \in V_1$ of MC-CAVI algorithm and for all $\varepsilon_1 > 0$, there exists a k_0 such that*

$$\lim_{N\to\infty} \mathrm{Prob}\left(|\mathscr{M}_N^k - \lambda^*| < \varepsilon_1 \text{ for some } k \le k_0\right) = 1.$$

The proofs of Lemma 1 and Theorem 1 can be found in Appendices A2 and A3, respectively.

3.2.4 Stopping Criterion and Sample Size

The method requires the specification of the Monte Carlo size N and a stopping rule.

Principled - but Impractical - Approach

As the algorithm approaches a local maximum, changes in ELBO should be getting closer to zero. To evaluate the performance of MC-CAVI, one could, in principle, attempt to monitor the evolution of ELBO during the algorithmic iterations. For current variational distribution $q = (q_1, \ldots, q_b)$, assume that MC-CAVI is about to update q_i with $q_i' = q_{i,N}'$, where the addition of the second subscript at this point emphasizes the dependence of the new value for q_i on the Monte Carlo size N. Define,

$$\Delta\mathrm{ELBO}(q,N) = \mathrm{ELBO}(q_{i-}, q_{i,N}', q_{i+}) - \mathrm{ELBO}(q).$$

If the algorithm is close to a local maximum, $\Delta\mathrm{ELBO}(q,N)$ should be close to zero, at least for sufficiently large N. Given such a choice of N, an MC-CAVI recursion should be terminated once $\Delta\mathrm{ELBO}(q,N)$ is smaller than a user-specified tolerance threshold. Assume that the random variable $\Delta\mathrm{ELBO}(q,N)$ has mean $\mu = \mu(q,N)$ and variance $\sigma^2 = \sigma^2(q,N)$. Chebychev's inequality implies that, with

probability greater than or equal to $(1-1/K^2)$, $\Delta\text{ELBO}(q,N)$ lies within the interval $(\mu - K\sigma, \mu + K\sigma)$, for any real $K > 0$. Assume that one fixes a large enough K. The choice of N and of a stopping criterion should be based on the requirements:

(i) $\sigma \leq \nu$, with ν a predetermined level of tolerance;

(ii) the effective range $(\mu - K\sigma, \mu + K\sigma)$ should include zero, implying that $\Delta\text{ELBO}(q,N)$ differs from zero by less than $2K\sigma$.

Requirement (i) provides a rule for the choice of N, which is assumed to be applied over all $1 \leq i \leq b$, for q in areas close to a maximiser, and requirement (ii) a rule for defining a stopping criterion. Unfortunately, the above considerations – based on the proper term $\text{ELBO}(q)$ that VI aims to maximise – involve quantities that are typically impossible to obtain analytically or via some reasonably expensive approximation.

Practical Considerations

Similarly to MCEM, it is recommended that N increases as the algorithm becomes more stable. It is computationally inefficient to start with a large value of N when the current variational distribution can be far from the maximiser. In practice, one may monitor the convergence of the algorithm by plotting relevant *statistics* of the variational distribution versus the number of iterations. We can declare that convergence has been reached when such traceplots show relatively small random fluctuations (due to the Monte Carlo variability) around a fixed value. At this point, one may terminate the algorithm or continue with a larger value of N, which will further decrease the traceplot variability. In the applications in this chapter, we typically have $N \leq 100$, so calculating, for instance, Effective Sample Sizes to monitor the mixing performance of the MCMC steps is not practical.

3.3 Numerical Examples – Simulation Study

In this section we illustrate MC-CAVI with two simulated examples. First, we apply MC-CAVI and CAVI on a simple model to highlight main features and implementation strategies. Then, we contrast MC-CAVI, MCMC, BBVI in a complex scenario with hard constraints.

3.3.1 Simulated Example 1

We generate $n = 10^3$ data points from $\mathrm{N}(10, 100)$ and fit the semi-conjugate Bayesian model

Example Model 1

$$\begin{aligned} x_1, \ldots, x_n &\sim \mathrm{N}(\vartheta, \tau^{-1}), \\ \vartheta &\sim \mathrm{N}(0, \tau^{-1}), \\ \tau &\sim \mathrm{Gamma}(1, 1). \end{aligned}$$

Let $\bar{x}$ be the data sample mean. In each iteration, the CAVI density function – see (1.14) – for τ is that of the Gamma distribution $\mathrm{Gamma}(\frac{n+3}{2}, \zeta)$, with

$$\zeta = 1 + \frac{(1+n)\mathbb{E}(\vartheta^2) - 2(n\bar{x})\mathbb{E}(\vartheta) + \sum_{j=1}^{n} x_j^2}{2},$$

whereas for ϑ that of the Normal distribution $\mathrm{N}(\frac{n\bar{x}}{1+n}, \frac{1}{(1+n)\mathbb{E}(\tau)})$. $(\mathbb{E}(\vartheta), \mathbb{E}(\vartheta^2))$ and $\mathbb{E}(\tau)$ denote the relevant expectations under the current CAVI distributions for ϑ and τ respectively; the former are initialized at 0 – there is no need to initialise $\mathbb{E}(\tau)$ in this case. Convergence of CAVI can be monitored, e.g., via the sequence of values of $\theta := (1+n)\mathbb{E}(\tau)$ and ζ. If the change in values of these two parameters is smaller than, say, 0.01%, we declare convergence. Figure 3.1 shows the traceplots of θ, ζ. Convergence is reached within 0.0017secs[1], after precisely two iterations, due to

[1] A Dell Latitude E5470 with Intel(R) Core(TM) i5-6300U CPU@2.40GHz is used for all experiments in this paper.

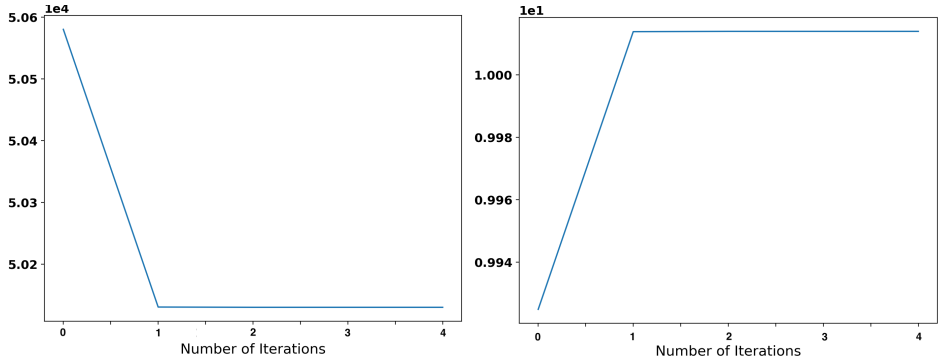

Figure 3.1: Tracplots of ζ (left), θ (right) from application of CAVI in Simulated Example 1.

the simplicity of the model. The resulting CAVI distribution for ϑ is $\mathrm{N}(9.6, 0.1)$, and for τ it is $\mathrm{Gamma}(501.5, 50130.3)$ so that $\mathbb{E}(\tau) \approx 0.01$.

Assume now that $q(\tau)$ is intractable. Since $\mathbb{E}(\tau)$ is required to update the approximate distribution of ϑ, an MCMC step can be employed to sample $\tau_1, \ldots, \tau_N$ from $q(\tau)$ to produce the Monte Carlo estimate $\widehat{\mathbb{E}}(\tau) = \sum_{j=1}^{N} \tau_j / N$. Within this MC-CAVI setting, $\widehat{\mathbb{E}}(\tau)$ will replace the exact $\mathbb{E}(\tau)$ during the algorithmic iterations. $(\mathbb{E}(\vartheta), \mathbb{E}(\vartheta^2))$ are initialised as in CAVI. For the first 10 iterations we set $N = 10$, and for the remaining ones, $N = 10^3$ to reduce variability. We monitor the values of $\widehat{\mathbb{E}}(\tau)$ shown in Figure 3.2. The figure shows that MC-CAVI has stabilized after about 15 iterations; algorithmic time was 0.0114secs. To remove some Monte Carlo variability, the final estimator of $\mathbb{E}(\tau)$ is produced by averaging the last 10 values of its traceplot, which gives $\widehat{\mathbb{E}}(\tau) = 0.01$, i.e. a value very close to the one obtained by CAVI. The estimated distribution of ϑ is $\mathrm{N}(9.6, 0.1)$, the same as with CAVI.

The performance of MC-CAVI depends critically on the choice N. Let A be the value of N in the burn-in period, B the number of burn-in iterations and C the value of N after burn-in. Figure 3.3 shows trace plots of $\widehat{\mathbb{E}}(\tau)$ under different settings of the triplet A-B-C.

As with MCEM, N should typically be set to a small number at the beginning of the iterations so that the algorithm can reach fast a region of relatively high prob-

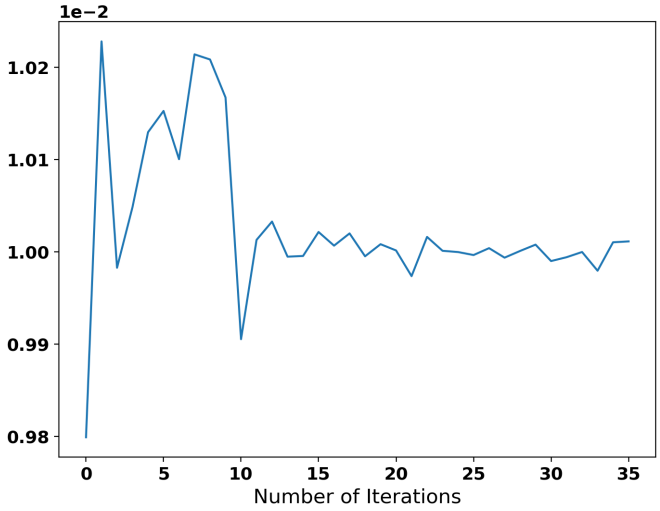

Figure 3.2: Traceplot of $\widehat{\mathbb{E}}(\tau)$ generated by MC-CAVI for Simulated Example 1, using $N = 10$ for the first 10 iterations of the algorithm, and $N = 10^3$ for the rest. The y-axis gives the values of $\widehat{\mathbb{E}}(\tau)$ across iterations.

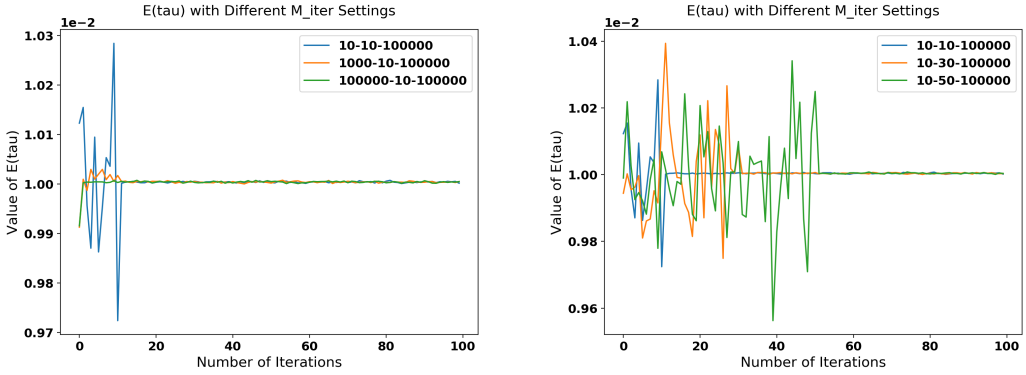

Figure 3.3: Traceplot of $\widehat{\mathbb{E}}(\tau)$ under different settings of A-B-C (respectively, the value of N in the burn-in period, the number of burn-in iterations and the value of N after burn-in) for Simulated Example 1.

A-B-C	10-10-10^5	10^3-10-10^5	10^5-10-10^5	10-30-10^5	10-50-10^5
time (secs)	0.4640	0.4772	0.5152	0.3573	0.2722
$\widehat{\mathbb{E}}(\tau)$	0.01	0.01	0.01	0.01	0.01

Table 3.1: Results of MC-CAVI for Simulated Example 1.

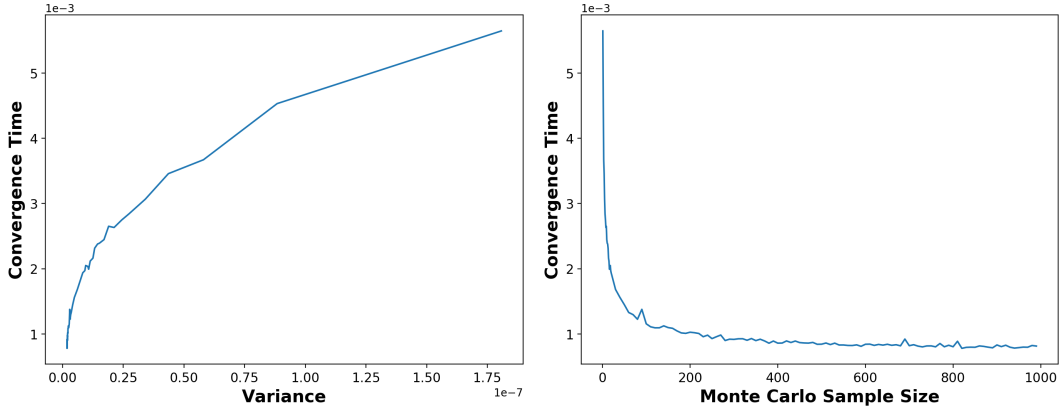

Figure 3.4: Plot of convergence time versus variance of $\widehat{\mathbb{E}}(\tau)$ (left panel) and versus Monte Carlo sample size N (right panel).

ability. N should then be increased to reduce algorithmic variability close to the convergence region. Figure 3.4 shows plots of convergence time versus variance of $\widehat{\mathbb{E}}(\tau)$ (left panel) and versus N (right panel). In VI, iterations are typically terminated when the (absolute) change in the monitored estimate is less than a small threshold. In MC-CAVI the estimate fluctuates around the limiting value after convergence. In the simulation in Figure 3.4, we terminate the iterations when the difference between the estimated mean (disregarding the first half of the chain) and the true value (0.01) is less than 10^{-5}. Figure 3.4 shows that: (i) convergence time decreases when the variance of $\widehat{\mathbb{E}}(\tau)$ decreases, as anticipated; (ii) convergence time decreases when N increases. In (ii), the decrease is most evident when N is still relatively small After N exceeds 200, convergence time remains almost fixed, as the benefit brought by the decrease of variance is offset by the cost of extra samples. (This is also in agreement with the practice of setting N equal to a small value at the initial iterations of the algorithm.)

3.3.2 Variance Reduction for BBVI

In non-trivial applications, the variability of the initial estimator $\nabla_\lambda \widehat{\mathrm{ELBO}}(q)$ within BBVI in (1.16) will typically be large, so variance reduction approaches such as Rao-Blackwellization and control variates [105] are also used. Rao-

Blackwellization [24] reduces variances by analytically calculating conditional expectations. In BBVI, within the factorization framework of (1.13), where $\lambda = (\lambda_1, \ldots, \lambda_b)$, and recalling identity (1.15) for the gradient, a Monte Carlo estimator for the gradient with respect to λ_i, $i \in \{1, \ldots, b\}$, can be simplified as

$$\nabla_{\lambda_i} \widehat{\mathrm{ELBO}}(q_i) = \frac{1}{N} \sum_{n=1}^{N} \left[\nabla_{\lambda_i} \log q_i(z_i^{(n)}|\lambda_i) \{ \log c_i(z_i^{(n)}, x) - \log q_i(z_i^{(n)}|\lambda_i) \} \right], \quad (3.4)$$

with $z_i^{(n)} \overset{iid}{\sim} q_i(z_i|\lambda_i)$, $1 \leq n \leq N$, and,

$$c_i(z_i, x) := \exp\left\{ \mathbb{E}_{-i}[\log p(z_{i_-}, z_i, z_{i_+}, x)] \right\}.$$

Depending on the model at hand, term $c_i(z_i, x)$ can be obtained analytically or via a double Monte Carlo procedure (for estimating $c_i(z_i^{(n)}, x)$, over all $1 \leq n \leq N$) – or a combination of thereof. In BBVI, control variates [113] can be defined on a per-component basis and be applied to the Rao-Blackwellized noisy gradients of ELBO in (3.4) to provide the estimator,

$$\nabla_{\lambda_i} \widehat{\mathrm{ELBO}}(q_i) = \frac{1}{N} \sum_{n=1}^{N} \left[\nabla_{\lambda_i} \log q_i(z_i^{(n)}|\lambda_i) \{ \log c_i(z_i^{(n)}, x) - \log q_i(z_i^{(n)}|\lambda_i) - \widehat{a}_i^* \} \right], \quad (3.5)$$

for the control,

$$\widehat{a}_i^* := \frac{\sum_{j=1}^{d_i} \widehat{\mathrm{Cov}}(f_{i,j}, g_{i,j})}{\sum_{j=1}^{d_i} \widehat{\mathrm{Var}}(g_{i,j})},$$

where $f_{i,j}$, $g_{i,j}$ denote the jth co-ordinate of the vector-valued functions f_i, g_i respectively, given below,

$$g_i(z_i) := \nabla_{\lambda_i} \log q_i(z_i|\lambda_i),$$
$$f_i(z_i) := \nabla_{\lambda_i} \log q_i(z_i|\lambda_i) \{ \log c_i(z_i, x) - \log q_i(z_i|\lambda_i) \}.$$

3.3.3 Simulated Example 2: Model with Hard Constraints

In this section, we discuss the performance and challenges of MC-CAVI, MCMC, BBVI for models where the support of the posterior – thus, also the variational distribution – involves hard constraints.

Here, we provide an example which offers a simplified version of the NMR problem discussed in Section 3.4 but allows for the implementation of BBVI, as the involved normalising constants can be easily computed. Moreover, as with other gradient-based methods, BBVI requires to tune the step-size sequence $\{\rho_k\}$ in (1.17), which might be a laborious task, in particular for increasing dimension. Although there are several proposals aimed to optimise the choice of $\{\rho_k\}$ (16, 77), MC-CAVI does not face such a tuning requirement.

We simulate data according to the following scheme: observations $\{y_j\}$ are generated from $\mathrm{N}(\vartheta+\kappa_j,\theta^{-1})$, $j=1,\dots,n$, with $\vartheta=6$, $\kappa_j=1.5\cdot\sin(-2\pi+4\pi(j-1)/n)$, $\theta=3$, $n=100$. We fit the following model:

Example Model 2

$$
\begin{aligned}
y_j \mid \vartheta,\kappa_j,\theta &\sim \mathrm{N}(\vartheta+\kappa_j,\theta^{-1}),\\
\vartheta &\sim \mathrm{N}(0,10),\\
\kappa_j \mid \psi_j &\sim \mathrm{TN}(0,10,-\psi_j,\psi_j),\\
\psi_j &\overset{i.i.d.}{\sim} \mathrm{TN}(0.05,10,0,2), \quad j=1,\dots,n,\\
\theta &\sim \mathrm{Gamma}(1,1).
\end{aligned}
$$

MCMC

We use a standard Metropolis-within-Gibbs. We set $y=(y_1,\dots,y_n)$, $\kappa=(\kappa_1,\dots,\kappa_n)$ and $\psi=(\psi_1,\dots,\psi_n)$. Notice that we have the full conditional dis-

tributions,

$$p(\vartheta|y,\theta,\kappa,\psi) = \mathrm{N}\big(\tfrac{\sum_{j=1}^{n}(y_j-\kappa_j)\theta}{\frac{1}{10}+n\theta}, \tfrac{1}{\frac{1}{10}+n\theta}\big),$$
$$p(\kappa_j|y,\theta,\vartheta,\psi) = \mathrm{TN}\big(\tfrac{(y_j-\vartheta)\theta}{\frac{1}{10}+\theta}, \tfrac{1}{\frac{1}{10}+\theta}, -\psi_j, \psi_j\big),$$
$$p(\theta|y,\vartheta,\kappa,\psi) = \mathrm{Gamma}\big(1+\tfrac{n}{2}, 1+\tfrac{\sum_{j=1}^{n}(y_j-\vartheta-\kappa_j)^2}{2}\big).$$

(Above, and in similar expressions written in the sequel, equality is meant to be properly understood as stating that 'the density on the left is equal to the density of the distribution on the right'.) For each ψ_j, $1 \le j \le n$, the full conditional is,

$$p(\psi_j|y,\theta,\vartheta,\kappa) \propto \frac{\phi\big(\frac{\psi_j-\frac{1}{20}}{\sqrt{10}}\big)}{\Phi\big(\frac{\psi_j}{\sqrt{10}}\big)-\Phi\big(\frac{-\psi_j}{\sqrt{10}}\big)}\,\mathbb{I}\big[|\kappa_j|<\psi_j<2\big], \quad j=1,\ldots,n,$$

where $\phi(\cdot)$ is the density of $\mathrm{N}(0,1)$ and $\Phi(\cdot)$ its cdf. The Metropolis-Hastings proposal for ψ_j is a Uniform variate from $\mathrm{U}(0,2)$.

MC-CAVI

For MC-CAVI, the logarithm of the joint distribution is given by,

$$\begin{aligned}\log p(y,\vartheta,\kappa,\psi,\theta) = const. + \tfrac{n}{2}\log\theta - \frac{\theta\sum_{j=1}^{n}(y_j-\vartheta-\kappa_j)^2}{2} - \frac{\vartheta^2}{2\cdot 10} \\ - \theta - \sum_{j=1}^{n}\frac{\kappa_j^2+(\psi_j-\frac{1}{20})^2}{2\cdot 10} \\ - \sum_{j=1}^{n}\log\big(\Phi\big(\tfrac{\psi_j}{\sqrt{10}}\big)-\Phi\big(\tfrac{-\psi_j}{\sqrt{10}}\big)\big),\end{aligned}$$

under the constraints,

$$|\kappa_j| < \psi_j < 2, \quad j=1,\ldots,n.$$

To comply with the above constraints, we factorise the variational distribution as,

$$q(\vartheta,\theta,\kappa,\psi) = q(\vartheta)q(\theta)\prod_{j=1}^{n} q(\kappa_j,\psi_j). \tag{3.6}$$

Here, for the relevant iteration k, we have,

$$q_k(\vartheta) = \mathrm{N}\big(\tfrac{\sum_{j=1}^{n}(y_j-\hat{\mathbb{E}}_{k-1}(\kappa_j))\mathbb{E}_{k-1}(\theta)}{\frac{1}{10}+n\mathbb{E}_{k-1}(\theta)}, \tfrac{1}{\frac{1}{10}+n\mathbb{E}_{k-1}(\theta)}\big),$$

$$q_k(\theta) = \mathrm{Gamma}\big(1+\tfrac{n}{2}, 1+\tfrac{\sum_{j=1}^{n}\mathbb{E}_{k,k-1}((y_j-\vartheta-\kappa_j)^2)}{2})\big),$$

$$q_k(\kappa_j,\psi_j) \propto \exp\big\{-\tfrac{\mathbb{E}_k(\theta)(\kappa_j-(y_j-\mathbb{E}_k(\vartheta)))^2}{2} - \tfrac{\kappa_j^2+(\psi_j-\frac{1}{20})^2}{2\cdot 10}\big\}\Big/\big(\Phi(\tfrac{\psi_j}{\sqrt{10}})-\Phi(\tfrac{-\psi_j}{\sqrt{10}})\big)$$
$$\cdot\,\mathbb{I}[\,|\kappa_j|<\psi_j<2], \qquad 1\le j\le n.$$

The quantity $\mathbb{E}_{k,k-1}((y_j-\vartheta-\kappa_j)^2)$ used in the second line above means that the expectation is considered under $\vartheta\sim q_k(\vartheta)$ and (independently) $\kappa_j\sim q_{k-1}(\kappa_j,\psi_j)$. Then, MC-CAVI develops as follows:

- Step 0: For $k=0$, initialize $\mathbb{E}_0(\theta)=1$, $\mathbb{E}_0(\vartheta)=4$, $\mathbb{E}_0(\vartheta^2)=17$.
- Step k: For $k\ge 1$, given $\mathbb{E}_{k-1}(\theta)$, $\mathbb{E}_{k-1}(\vartheta)$, execute:
  - For $j=1,\ldots,n$, apply an MCMC algorithm – with invariant law $q_{k-1}(\kappa_j,\psi_j)$ – consisted of a number, N, of Metropolis-within-Gibbs iterations carried out over the relevant full conditionals,

$$q_{k-1}(\psi_j|\kappa_j) \propto \frac{\phi(\frac{\psi_j-\frac{1}{20}}{\sqrt{10}})}{\Phi(\frac{\psi_j}{\sqrt{10}})-\Phi(\frac{-\psi_j}{\sqrt{10}})}\,\mathbb{I}[\,|\kappa_j|<\psi_j<2],$$

$$q_{k-1}(\kappa_j|\psi_j) = \mathrm{TN}\big(\tfrac{(y_j-\mathbb{E}_{k-1}(\vartheta))\mathbb{E}_{k-1}(\theta)}{\frac{1}{10}+\mathbb{E}_{k-1}(\theta)}, \tfrac{1}{\frac{1}{10}+\mathbb{E}_{k-1}(\theta)}, -\psi_j, \psi_j\big).$$

As with the full conditional $p(\psi_j|y,\theta,\vartheta,\kappa)$ within the MCMC sampler, we use a Uniform proposal $\mathrm{U}(0,2)$ at the Metropolis-Hastings step applied for $q_{k-1}(\psi_j|\kappa_j)$. For each k, the N iterations begin from the (κ_j,ψ_j)-values obtained at the end of the corresponding MCMC iterations at step $k-1$, with very first initial values being $(\kappa,\psi_j)=(0,1)$. Use the N samples to obtain $\hat{\mathbb{E}}_{k-1}(\kappa_j)$ and $\hat{\mathbb{E}}_{k-1}(\kappa_j^2)$.

– Update the variational distribution for ϑ,

$$q_k(\vartheta)=\mathrm{N}\big(\tfrac{\sum_{j=i}^{n}(y_j-\hat{\mathbb{E}}_{k-1}(\kappa_j))\mathbb{E}_{k-1}(\theta)}{\frac{1}{10}+n\mathbb{E}_{k-1}(\theta)},\tfrac{1}{\frac{1}{10}+n\mathbb{E}_{k-1}(\theta)}\big)$$

and evaluate $\mathbb{E}_k(\vartheta)$, $\mathbb{E}_k(\vartheta^2)$.

– Update the variational distribution for θ,

$$q_k(\theta)=\mathrm{Gamma}\big(1+\tfrac{n}{2},1+\tfrac{\sum_{j=1}^{n}\hat{\mathbb{E}}_{k,k-1}((y_j-\vartheta-\kappa_j)^2)}{2}\big)$$

and evaluate $\mathbb{E}_k(\theta)$.

- Iterate until convergence.

BBVI

For BBVI we assume a variational distribution $q(\theta,\vartheta,\kappa,\psi\,|\,\boldsymbol{\alpha},\boldsymbol{\gamma})$ that factorises as in the case of CAVI in (3.6), where

$$\boldsymbol{\alpha}=(\alpha_\vartheta,\alpha_\theta,\alpha_{\kappa_1},\ldots,\alpha_{\kappa_n},\alpha_{\psi_1},\ldots,\alpha_{\psi_n})\,,$$

$$\boldsymbol{\gamma}=(\gamma_\vartheta,\gamma_\theta,\gamma_{\kappa_1},\ldots,\gamma_{\kappa_n},\gamma_{\psi_1},\ldots,\gamma_{\psi_n})$$

to be the variational parameters. Individual marginal distributions are chosen to agree – in type – with the model priors. In particular, we set,

$$q(\vartheta) = \mathrm{N}(\alpha_\vartheta, \exp(\gamma_\vartheta)),$$

$$q(\theta) = \mathrm{Gamma}(\exp(\alpha_\theta), \exp(\gamma_\theta)),$$

$$q(\kappa_j, \psi_j) = \mathrm{TN}(\alpha_{\kappa_j}, \exp(2\gamma_{\kappa_j}), -\psi_j, \psi_j) \otimes \mathrm{TN}(\alpha_{\psi_j}, \exp(2\gamma_{\psi_j}), 0, 2), \quad 1 \le j \le n.$$

It is straightforward to derive the required gradients (see Appendix A4 for the analytical expressions). BBVI is applied using Rao-Blackwellization and control variates for variance reduction. The algorithm is as follows,

- Step 0: Set $\eta = 0.5$; initialise $\boldsymbol{\alpha}^0 = 0$, $\boldsymbol{\gamma}^0 = 0$ with the exception $\alpha_\vartheta^0 = 4$.
- Step k: For $k \ge 1$, given $\boldsymbol{\alpha}^{k-1}$ and $\boldsymbol{\gamma}^{k-1}$ execute:
  - Draw $(\vartheta^i, \theta^i, \kappa^i, \psi^i)$, for $1 \le i \le N$, from $q_{k-1}(\vartheta)$, $q_{k-1}(\theta)$, $q_{k-1}(\kappa, \psi)$.
  - With the samples, use (3.5) to evaluate:

$$\begin{aligned}
&\nabla^k_{\alpha_\vartheta} \widehat{\mathrm{ELBO}}(q(\vartheta)), \quad \nabla^k_{\gamma_\vartheta} \widehat{\mathrm{ELBO}}(q(\vartheta)),\\
&\nabla^k_{\alpha_\theta} \widehat{\mathrm{ELBO}}(q(\theta)), \quad \nabla^k_{\gamma_\theta} \widehat{\mathrm{ELBO}}(q(\theta)),\\
&\nabla^k_{\alpha_{\kappa_j}} \widehat{\mathrm{ELBO}}(q(\kappa_j, \psi_j)), \quad \nabla^k_{\gamma_{\kappa_j}} \widehat{\mathrm{ELBO}}(q(\kappa_j, \psi_j)), \quad 1 \le j \le n,\\
&\nabla^k_{\alpha_{\psi_j}} \widehat{\mathrm{ELBO}}(q(\kappa_j, \psi_j)), \quad \nabla^k_{\gamma_{\psi_j}} \widehat{\mathrm{ELBO}}(q(\kappa_j, \psi_j)), \quad 1 \le j \le n.
\end{aligned}$$

    (Here, superscript k at the gradient symbol ∇ specifies the BBVI iteration.)

  - Evaluate $\boldsymbol{\alpha}^k$ and $\boldsymbol{\gamma}^k$:

$$(\boldsymbol{\alpha}, \boldsymbol{\gamma})^k = (\boldsymbol{\alpha}, \boldsymbol{\gamma})^{k-1} + \rho_k \nabla^k_{(\boldsymbol{\alpha}, \boldsymbol{\gamma})} \widehat{\mathrm{ELBO}}(q),$$

where $q = (q(\vartheta), q(\theta), q(\kappa_1, \psi_1), \ldots, q(\kappa_n, \psi_n))$. For the learning rate, we employed the AdaGrad algorithm [39] and set $\rho_k = \eta \, \mathrm{diag}(G_k)^{-1/2}$, where G_k is a matrix equal to the sum of the first k iterations of the outer products of the gradient, and $\mathrm{diag}(\cdot)$ maps a matrix to its diagonal version.

- Iterate until convergence.

Results

The three algorithms have different stopping criteria. We run each for 100secs for a fair comparison. Please note that with 100secs, MCMC and MC-CAVI have already achieved satisfying convergence (for MCMC, the z-scores of Geweke's Diagnostic for ϑ and θ are 0.19 and 0.28 respectively; for MC-CAVI, the chain fluctuates stably around a small area) while BBVI does not. A summary of results is given in Table 3.2. Model fitting plots and algorithmic traceplots are shown in Figure 3.5.

	MCMC	MC-CAVI	BBVI
Iterations	No. Iterations = 2,500 Burn-in = 1,250	No. Iterations = 300 $N = 10$ Burn-in = 150	No. Iterations = 100 $N = 10$
ϑ	5.927 (0.117)	5.951 (0.009)	6.083 (0.476)
θ	1.248 (0.272)	8.880 (0.515)	0.442 (0.172)

Table 3.2: Summary of results: last two rows show the average for the corresponding parameter (in horizontal direction) and algorithm (in vertical direction), after burn-in (the number in brackets is the corresponding standard deviation). All algorithms were executed for 10^2secs. The first row gives some algorithmic details.

Table 3.2 indicates that all three algorithms approximate the posterior mean of ϑ effectively; the estimate from MC-CAVI has smaller variability than the one of BBVI; the opposite holds for the variability in the estimates for θ. Figure 3.5 shows that the traceplots for BBVI are unstable, a sign that the gradient estimates have high variability. In contrast, MCMC and MC-CAVI perform rather well. Figure 3.6 shows the 'true' posterior density of ϑ (obtained from an expensive MCMC with

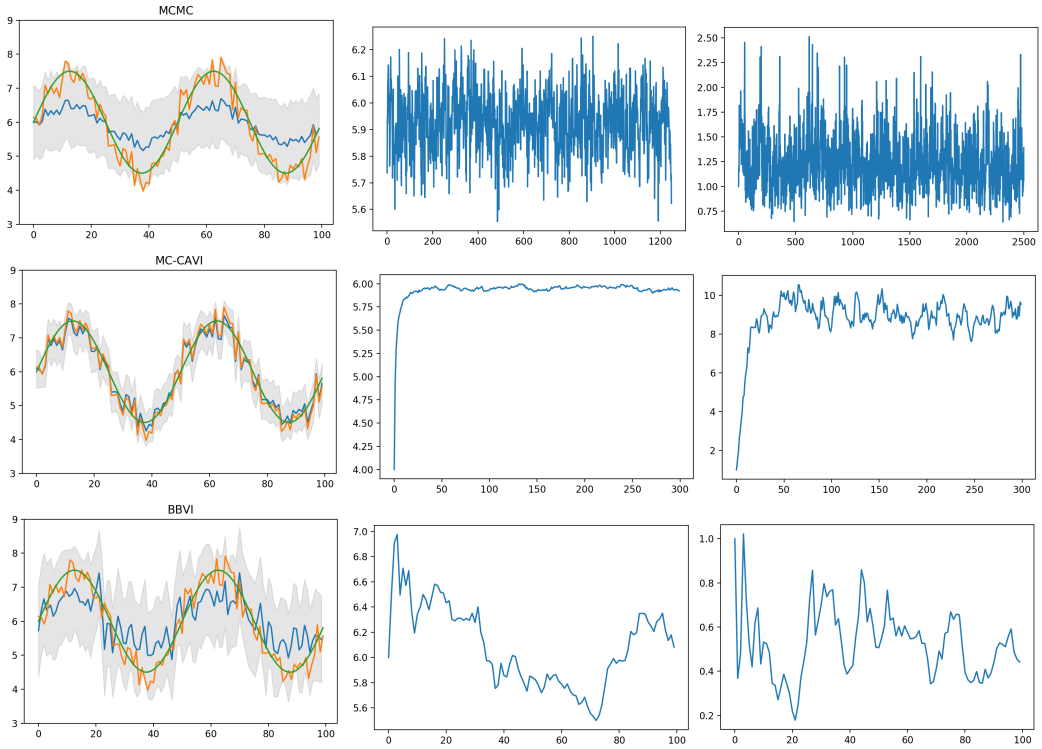

Figure 3.5: Model fit (left panel), traceplots of ϑ (middle panel) and traceplots of θ (right panel) for the three algorithms: MCMC (first row), MC-CAVI (second row) and BBVI (third row) – for Example Model 2 – when allowed 100secs of execution. In the plots showing model fit, the green line represents the data without noise, the orange line the data with noise; the blue line shows the corresponding posterior means and the grey area the pointwise 95% posterior credible intervals.

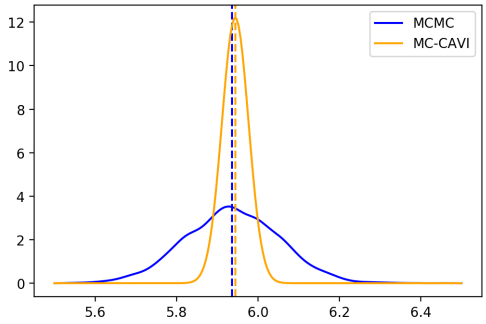

Figure 3.6: Density plots for the true posterior of ϑ (blue line) – obtained via an expensive MCMC – and the corresponding approximate distribution provided by MC-CAVI.

10,000 iterations – 5,000 burn-in) and the corresponding approximation obtained via MC-CAVI. In this case, the variational approximation is quite accurate at the

estimation of the mean but underestimates the posterior variance (rather typically for a VI method). We mention that for BBVI we also tried to use normal laws as variational distributions – as this is mainly the standard choice in the literature – however, in this case, the performance of BBVI deteriorated even further.

3.4 Application to ^{1}H NMR Spectroscopy

We demonstrate the utility of MC-CAVI in a statistical model proposed in the field of metabolomics by Astle et al. [3], and used in NMR (Nuclear Magnetic Resonance) data analysis (Section 1.6). The aim of the analysis is: (i) to deconvolve resonance peak in the spectrum and assign them to a particular metabolite; (ii) estimate the abundance of the catalogued metabolites; (iii) model the component of a spectrum that cannot be assigned to known compounds.

BATMAN is an R package for estimating metabolite concentrations from NMR spectral data using a specifically designed MCMC algorithm [60] to perform posterior inference from the Bayesian model described in Section 1.6. We implement a MC-CAVI version of BATMAN and compare its performance with the original MCMC algorithm. Details of the implementation of MC-CAVI are given in the Appendix. Due to the complexity of the model and the datasize, it is challenging for both algorithms to reach convergence. We run the two methods, MC-CAVI and MCMC, for approximately an equal amount of time, to analyse a full spectrum with 1,530 data points and modelling parametrically 10 metabolites. We fix the number of iterations for MC-CAVI to 1,000, with a burn-in of 500 iterations; we set the Monte Carlo size to $N = 10$ for all iterations. The execution time for this MC-CAVI algorithms is $2,048$secs. For the MCMC algorithm, we fix the number of iterations to 2,000, with a burn-in of 1,000 iterations. This MCMC algorithm has an execution time of $2,098$secs.

In ^{1}H NMR analysis, β (the concentration of metabolites in the biofluid) and $\delta^*_{m,u}$

(the peak positions) are the most important parameters from a scientific point of view. Traceplots of four examples (β_3, β_4, β_9 and $\delta_{4,1}$) are shown in Figure 3.7. These four parameters are chosen due to the different performance of the two methods, which are closely examined in Figure 3.9. For β_3 and β_9, traceplots are still far from convergence for MCMC, while they move towards the correct direction (see Figure 3.7) when using MC-CAVI. For β_4 and $\delta_{4,1}$, both parameters reach a stable regime very quickly in MC-CAVI, whereas the same parameters only make local moves when implementing MCMC. For the remaining parameters in the model, both algorithms present similar results. Please note that, due to the difficulty involved in convergence of both algorithms, two algorithms are compared with same running time. The convergence of MC-CAVI is not better than that of MCMC in terms of convergence diagnostics. The argument we claim that is that, although both algorithms have not reached convergence, according to the close examination we performed with Figure 3.9 later, with the same running time, MC-CAVI algorithms reached an area more "correct" than MCMC.

Figure 3.8 shows the fit obtained from both the algorithms, while Table 3.3 reports posterior estimates for β. From Figure 3.8, it is evident that the overall performance of MC-CAVI is similar as that of MCMC since in most areas, the metabolite fit (orange line) captures the shape of the original spectrum quite well. Table 3.3 shows that, similar to standard VI behaviour, MC-CAVI underestimates the variance of the posterior density. We examine in more detail the posterior distribution of the β coefficients for which the posterior means obtained with the two algorithms differ more than 1.0e-4. Figure 3.9 shows that MC-CAVI manages to capture the shapes of the peaks while MCMC does not, around ppm values of 2.14 and 3.78, which correspond to spectral regions where many peaks overlap making peak deconvolution challenging. This is probably due to the faster convergence of MC-CAVI. Figure 3.9 shows that for areas with no overlapping (e.g. around ppm values of 2.66 and 7.53), MC-CAVI and MCMC produce similar results.

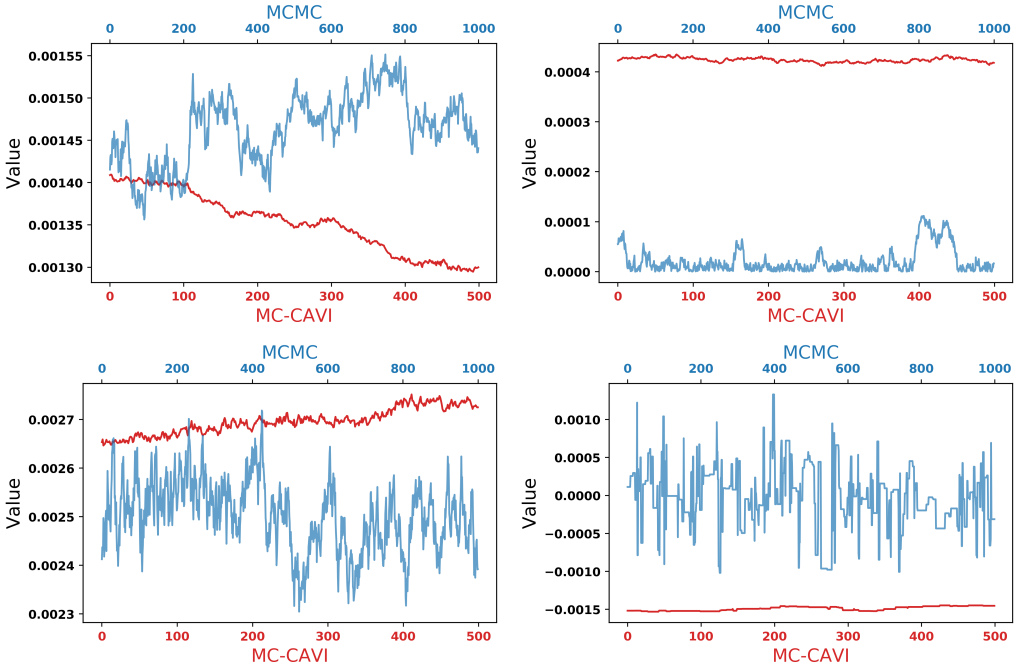

Figure 3.7: Traceplots of Parameter Value against Number of Iterations after the burn-in period for β_3 (upper left panel), β_4 (upper right panel), β_9 (lower left panel) and $\delta_{4,1}$ (lower right panel). The y-axis corresponds to the obtained parameter values (the mean of the distribution q for MC-CAVI and traceplots for MCMC). The red line shows the results from MC-CAVI and the blue line from MCMC. Both algorithms are executed for the same (approximately) amount of time.

		β_1	β_2	$\boldsymbol{\beta_3}$	$\boldsymbol{\beta_4}$	β_5
MC-CAVI	mean	6.0e-6	7.8e-5	1.4e-3	4.2e-4	2.6e-5
	sd	1.8e-11	4.0e-11	1.3e-11	1.0e-11	6.2e-11
MCMC	mean	1.2e-5	4.0e-5	1.5e-3	2.1e-5	3.4e-5
	sd	1.1e-10	5.0e-10	1.6e-9	6.4e-10	3.9e-10
		β_6	β_7	β_8	$\boldsymbol{\beta_9}$	β_{10}
MC-CAVI	mean	6.1e-4	3.0e-5	1.9e-4	2.7e-3	1.0e-3
	sd	1.5e-11	1.6e-11	3.9e-11	1.6e-11	3.6e-11
MCMC	mean	6.0e-4	3.0e-5	1.8e-4	2.5e-3	1.0e-3
	sd	2.3e-10	7.5e-11	3.7e-10	5.1e-9	7.9e-10

Table 3.3: Estimation of β obtained with MC-CAVI and MCMC. (The coefficients of β for which the posterior means obtained with the two algorithms differ by more than 1.0e-4 are shown in bold.)

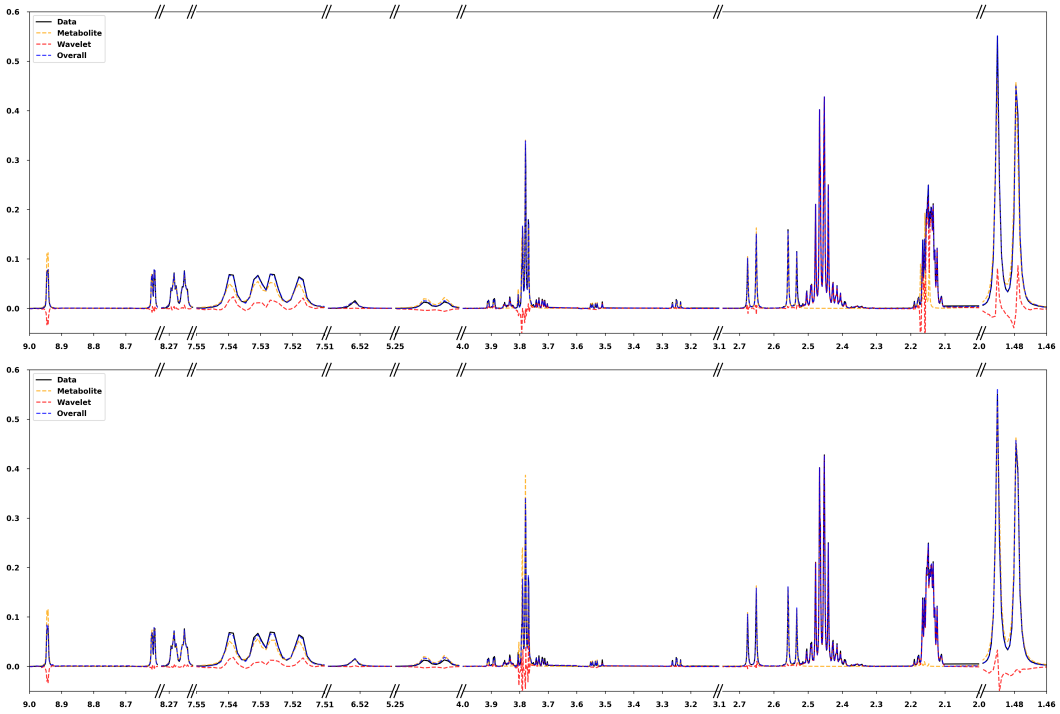

Figure 3.8: Comparison of MC-CAVI and MCMC in terms of Spectral Fit. The upper panel shows the Spectral Fit from MC-CAVI algorithm. The lower panel shows the Spectral Fit from MCMC algorithm. The x-axis corresponds to chemical shift measure in ppm. The y-axis corresponds to standard density.

Comparing MC-CAVI and MCMC's performance in the case of the NMR model, we can draw the following conclusions:

- In NMR analysis, if many peaks overlap (see Figure 3.9), MC-CAVI can provide better results than MCMC.
- In high-dimensional models, where the number of parameters grows with the size of data, MC-CAVI can converge faster than MCMC.
- Choice of N is important for optimising the performance of MC-CAVI. Building on results derived for other Monte Carlo methods (e.g. MCEM), it is reasonable to choose a relatively small number of Monte Carlo iterations at the beginning when the algorithm can be far from regions of parameter space of high posterior probability, and gradually increase the number of Monte Carlo

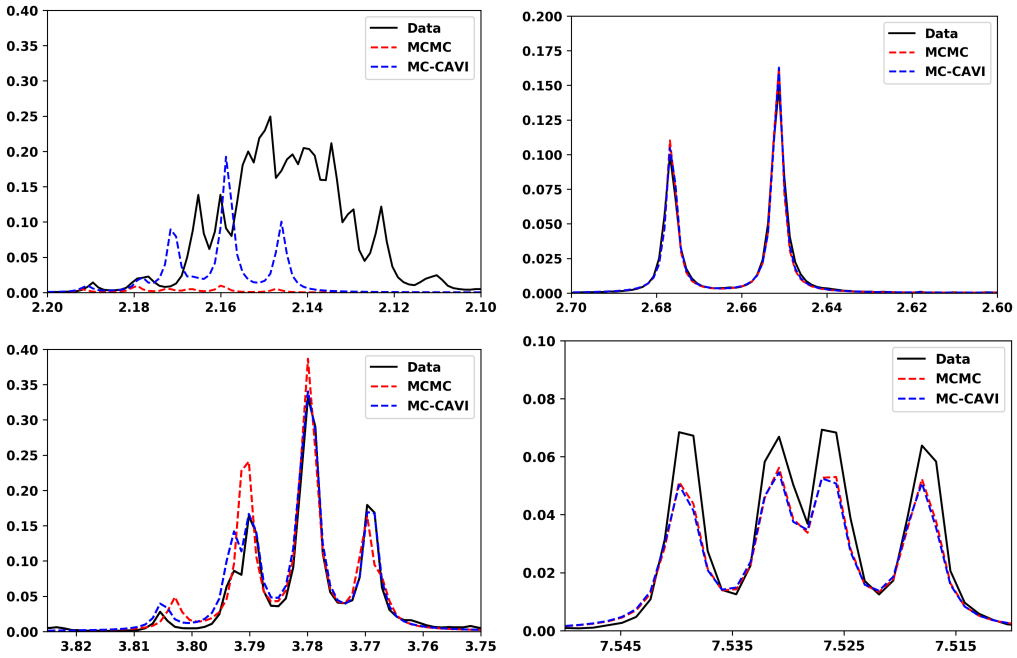

Figure 3.9: Comparison of Metabolites Fit obtained with MC-CAVI and MCMC. The x-axis corresponds to chemical shift measure in ppm. The y-axis corresponds to standard density. The upper left panel shows areas around ppm value 2.14 (β_4 and β_9). The upper right panel shows areas around ppm 2.66 (β_6). The lower left panel shows areas around ppm value 3.78 (β_3 and β_9). The lower right panel shows areas around ppm 7.53 (β_{10}).

iterations, with the maximum number taken once the algorithm has reached a mode.

3.5 Conclusion

As a combination of VI and MCMC, MC-CAVI has the potential to improve NMR spectroscopy analysis and provides a powerful inferential tool particularly in high dimensional settings when full posterior inference is computationally demanding and the application of optimization and of noisy-gradient-based approaches, e.g. BBVI, is hindered by the presence of hard constraints. The MCMC step of MC-CAVI is necessary to deal with parameters for which VI approximation distributions are difficult or impossible to derive, for example due to the impossibility to

derive closed-form expression for the normalising constant. General Monte Carlo algorithms such as sequential Monte Carlo and Hamiltonian Monte Carlo can be incorporated within MC-CAVI. Compared with MCMC, the VI step of MC-CAVI speeds up convergence and provides reliable estimates in a shorter time. Moreover, MC-CAVI scales better in high-dimensional settings. As an optimization algorithm, MC-CAVI's convergence monitoring is easier than MCMC. Moreover, MC-CAVI offers a flexible alternative to BBVI. This latter algorithm, although very general and suitable for a large range of complex models, depends crucially on the quality of the approximation to the true target provided by the variational distribution, which in high dimensional setting (in particular with hard constraints) is very difficult to assess.

Chapter 4

Bayesian deconvolution and quantification of metabolites from *J*-resolved NMR spectroscopy

4.1 Background

While metabolite identification and quantification in 1D NMR spectroscopy are severely hindered by resonance overlapping, JRES, a popular 2D method for metabolomics, disperses the overlapping resonances into a second dimension. With this extra dimension, JRES has the potential to significantly reduce congestion, and enhance metabolite identification and estimation [84].

Standard analysis of JRES data is often based on 1D projections of the 2D spectra. For example, Viant [130] perform multivariate statistical analyses for JRES metabolomics data by taking projections of each 2D spectrum onto the chemical shift axis. For instance, 1D projections of JRES spectra inevitably discard the spin-spin coupling measurements, which potentially become important for further discrimination between different metabolites, especially within complex biological samples. *J*-coupling also has the advantage that the coupling patterns are less sen-

sitive to changes in pH than chemical shift values [93]. This is not an ideal strategy as it does not allow to fully exploit information from one dimension.

Gómez et al. [55] combine 2D JRES with 1D NMR spectra to avoid peak misidentification. Their quantification step, however, is still performed on the 1D spectrum. Kikuchi et al. [76] construct a database for 2D JRES spectra from 598 metabolite standards and develop analytic tools for absolute quantification. However, their quantification tool only supports 38 commonly observed major metabolites. Another typical approach is to unfold the 2D data into a single row vector which can then be used in supervised or unsupervised machine learning algorithms. For example, Parsons et al. [100] are able to discriminate liver samples from fish derived from different polluted rivers using this simple approach. Again, this process does not make full use of the information provided by the second dimension.

The most widely used statistical methods to analyse 2D JRES data from their original format are: (i) binning the spectrum to reduce dimensionality and evaluating summary statistics; (ii) unsupervised multivariate clustering techniques, such as Ward's algorithm or K-means, applied to bucketed or original spectral data; and (iii) peak alignment followed by pattern recognition methods using principal component analysis or partial least squares regression. The limitations of bucketing spectral data are well documented [30, 49] and, in general, none of these methods fully exploit the information in the spectrum. While these methods usually lead to the identification of spectral regions associated, for example, to a phenotype of interest, they still require extensive work for the identification and estimation of concentration of metabolites. Perhaps, even more importantly, they do not provide measures of uncertainty associated with the estimates.

Potentially the most accurate approach to analyze an intact 2D JRES spectrum is fitting manually each individual resonance to the theoretical peak shape of a certain metabolite. Peak identification is complicated by variations in peak positions

between spectra, caused by inevitable and uncontrollable changes in experimental conditions and differences in the chemical properties of the biological samples. Expert spectroscopist deconvolution is rarely practical for JRES spectra because it is time consuming and requires knowledge about metabolite resonance patterns. Targeted profiling [135], usually performed in 1D against a standard library of metabolite resonance peaks, reduces the requirement of expert spectroscopist knowledge but is still labour intensive. Therefore, we aim to develop a full likelihood based approach to analyse 2D JRES data, which allows for expert guided automatic deconvolution, identification and quantification of metabolites.

Contribution of this section: Since JRES datasets are large (typically $50-100$ times larger than comparable 1D NMR spectra) and heavily structured, specialized models and appropriate tools are required to perform metabolite quantification. To the best of our knowledge, there are no efficient statistical methods available for analysing JRES spectra, which automatically combine the data-generating mechanisms and the extensive prior knowledge available in online databases, and at the same time provide measures of uncertainty. In this section, we develop a fully likelihood based approach to analyse 2D JRES data from complex biological mixtures, which allows for expert guided automatic deconvolution, identification and quantification of metabolites. The advantages of our method are that it allows direct quantification of metabolites drawn from a library of known compounds, disambiguation of assignment of highly overlapping resonances, deconvolution of signals in highly crowded regions, and estimates of uncertainty in relative concentrations and peak positions. Note that in many applications only relative concentration estimation, i.e. estimation of the ratio of concentrations between samples, is feasible since absolute quantification usually involves calibration of signals from a biological mixture of interest using reference signals from a standard containing a detectable compound of unknown concentration.

Our approach is based on a combination of theoretical templates and B-spline tight wavelet frames. The incorporation of theoretical or empirical metabolite templates is a clear advantage in terms of model interpretability as compared to common analysis tools in metabolomics such as bucketing, principal component analysis and partial least squares or to a model based only on basis function representation of the spectrum. We perform posterior inference through specially devised Markov chain Monte Carlo (MCMC) methods. Finally, we demonstrate the effectiveness of our approach on simulated data and via analyses of datasets from serum and urine.

4.2 Modelling

Acquisition of NMR data requires sampling at regularly spaced time points to yield time domain data, which needs to be transformed to Fourier/frequency domain (as shown in Figure 1.3). The Fourier transform is necessary to convert the spectrum represented by a series of cosines in time domain to an easily recognisable spectrum in frequency domain. Next, the resulting 2D frequency spectra require specific processing, which comprises mainly of two steps: tilting the spectrum, followed by symmetrisation. Tilting involves moving the centre of the peaks corresponding to the same multiplet in the J-coupling dimension so that they are aligned in the chemical shift dimension. Points other than the centre are also moved in a similar manner. In other words, after tilting, peak maxima in each multiplet appear at the same resonance frequency. Since the tilted peaks have now been subjected to a shearing transformation, the resultant peak shapes have changed from the initial unprocessed spectrum. Consequently, the spectrum has to be symmetrised, forcing the signal intensities to become symmetric around the centre line of the spectrum along the J-coupling dimension. After symmetrisation, the peaks are truncated, but still centred. eAfter this standard preprocessing, which is typically performed fully automatically with the spectrometer manufacturer's proprietary software (or using

publicly available packages such as NMRglue [62], a frequency-domain 2D JRES spectrum, as exemplified in Figure 1.3, is given by position vectors $\mathbf{x} = (x_1, \ldots, x_{N_C})$ on the chemical shift axis and $\mathbf{y} = (y_1, \ldots, y_{N_J})$ on the J-coupling axis, together with a measurement matrix $\mathbf{z} = (z_{ij})_{i=1,\ldots,N_C; j=1,\ldots,N_J}$ whose elements are the resonance intensities at the usually uniformly spaced positions (x_i, y_j). Depending on the resolution of the spectrum and the size of the region under consideration, N_C typically is of the order $10^3 - 10^4$, while N_J typically is of the order $10^2 - 10^3$. The intensity measurements are corrupted by noise and therefore, although inherently positive quantities, may in some cases be negative valued. We standardize the intensities to satisfy $\sum_{i,j} z_{ij} = 1$.

We model $\mathbf{z} \mid \mathbf{x}, \mathbf{y}$ assuming that $z_{ij} \mid \mathbf{x}, \mathbf{y}$ are independent Normal random variables with

$$\mathbb{E}(z_{ij} \mid \mathbf{x}, \mathbf{y}) = \phi(x_i, y_j) + \xi(x_i, y_j), \quad \text{for } 1 \leq i \leq N_C \text{ and } 1 \leq j \leq N_J. \tag{4.1}$$

The ϕ component of the model corresponds to signal from targeted metabolites which we aim to quantify and for which prior information in the form of spectral signatures is available, either catalogued in public databases or through expert knowledge. The ξ component of the model represents the signal generated by untargeted and/or unknown metabolites or other molecules and may, if necessary, include partial signals from metabolites whose residual resonances are modeled in the ϕ component. This construction mirrors an equivalent modelling strategy developed by Astle et al. [3] for 1D NMR data. We model the ϕ component parametrically via continuous functions of continuous chemical shift and J-coupling information, using the physical theory of J-resolved NMR [see, e.g., 84]. The ξ component is modelled nonparametrically using a wavelet system constructed from a piecewise linear B-spline [see 38].

4.2.1 Modelling of catalogued metabolite signal

In theory, resonance signatures of different metabolites are independent and aggregate in the JRES spectrum by convolution, with an intensity proportional to molecular abundance. Each molecular compound has a specific spectral signature given by a set of multiplets across the spectrum. These multiplets are characterized by their position δ on the chemical shift axis and the position ζ of their individual peaks on the J-coupling axis.

More precisely, the targeted signal is a linear combination of the signatures of M different targeted metabolites, i.e.

$$\phi(\delta,\zeta) = \sum_{m=1}^{M} \beta_m t_m(\delta,\zeta) \quad \text{for } (\delta,\zeta) \in \mathbb{R}^2, \tag{4.2}$$

where the t_m are continuous template functions specifying the JRES signatures of the metabolites, with concentrations β_m that are proportional to the molecular abundance of the m-th metabolite in the biological mixture. The number of targeted metabolites M is specified by the researcher and depends on the available prior information and the scientific problem. In general, M varies between one to several hundreds.

The JRES signatures t_m of the metabolites are a superposition of multiplets, each of which is in turn a superposition of individual peaks. Multiplets appear at certain positions on the chemical shift and J-coupling axes. The number of peaks, their distances from each other and relative heights can be used for metabolite identification. More precisely,

$$t_m(\delta,\zeta) = \sum_{u} \rho_{mu} g_{mu}(\delta - \delta^{\star}_{mu}, \zeta), \tag{4.3}$$

where u is indexing the multiplets g_{mu} belonging to the m-th metabolite. The chemical shift parameter $\delta^{\star}_{mu}$ of the multiplet specifies the position of the centre

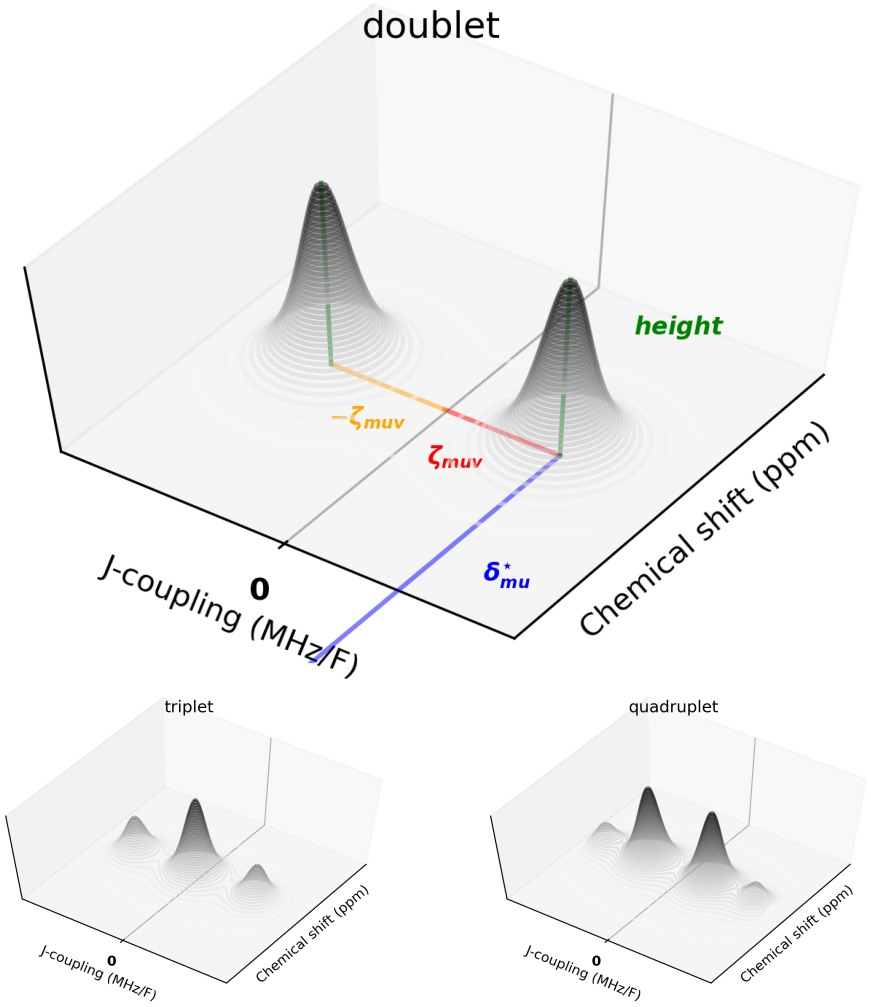

Figure 4.1: Peak configurations of some common multiplet types. The x-axis indicates chemical shift while the y-axis indicates the J-coupling. The upper panel shows a doublet with chemical shift $\delta^{\star}_{mu}$ and peak offset ζ_{muv}. The lower panel shows a triplet and quadruplet.

of mass of g_{mu}. The coefficients ρ_{mu} are usually equal to the number of protons in a molecule of the metabolite that contributes resonance signal to the u-th multiplet. Due to relaxation effects [67] the ρ_{mu} may not always be positive integers, in which case they have to be interpreted as “effective" proton contributions. The volume $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_{mu}(\delta,\zeta)\,d\delta\,d\zeta$ is assumed to be constant over m and u. Thus the volume under each t_m is proportional to the number $\sum_u \rho_{mu}$ of resonating protons in the m-th molecule, giving a measure of abundance. These observations will become crucial when we describe our shrinkage strategy in Section 4.3.

Besides few exceptions, the peak configurations of the multiplets g_{mu} can be classified into several common types, such as doublets, triplets, or doublet of doublets (see Figure 4.1). This classification, together with a small number of continuous quantities called J-coupling constants, which determine the distance of each peak from the centre of the multiplet along the J-coupling axis, completely parametrize a multiplet. We model multiplets g_{mu} as weighted averages of V_{mu} translated generalized bivariate Student-t densities $f_{\sigma_1\sigma_2\nu}$ with zero mean and zero correlation, which we will discuss in more detail in (4.5) below. More precisely,

$$g_{mu}(\delta,\zeta) = \sum_{v=1}^{V_{mu}} w_{muv} f_{\sigma_1\sigma_2\nu}(\delta, \zeta - \zeta_{muv}), \tag{4.4}$$

where the weights w_{muv} (which over v sum to one, and are available through data banks and expert knowledge) determine the relative heights of the peaks in the multiplet. The translation parameters ζ_{muv} determine the J-coupling offsets of the peaks from the centre of mass of the multiplet. Multiplets are usually symmetric around $\zeta = 0$, with $\{-\zeta_{muv}\}_{v=1,\ldots,V_{mu}} = \{\zeta_{muv}\}_{v=1,\ldots,V_{mu}}$, and $w_{muv'} = w_{muv}$ whenever $\zeta_{muv'} = -\zeta_{muv}$, see Figure 4.1.

Under ideal experimental conditions, the individual peaks in 1D NMR spectra have the shape of Lorentzians [25]. In 2D JRES spectra the tensor product of two Lorentzian curves may be used to fit individual peaks, however, the precise mathematical description of peak shapes in JRES spectra has yet to be determined [54]. In many types of spectroscopy, Voigt profiles are used to model peak shapes [19]. They can be understood as a convolution of Lorentzian and Gaussian profiles, each of which is derived from different underlying physical processes. However, the relative importance of these processes is difficult to estimate from the data and is usually inferred from evidence for light/heavy tails. We therefore choose to model peaks by generalized bivariate Student-t distribution kernels with zero mean and

zero correlation given by

$$f_{\sigma_1\sigma_2\nu}(\delta,\zeta-\zeta_{muv}) = \frac{\Gamma((\nu+2)/2)}{\Gamma(\nu/2)\pi\nu\sigma_1\sigma_2}\left(1+\frac{1}{\nu}\left(\frac{\delta^2}{\sigma_1^2}+\frac{(\zeta-\zeta_{muv})^2}{\sigma_2^2}\right)\right)^{-(\nu+2)/2} \quad \text{for } (\delta,(\zeta-\zeta_{muv}))\in\mathbb{R}^2, \tag{4.5}$$

where σ_1,σ_2 are scaling parameters controlling peak width, ν represents the number of degrees of freedom controlling the tail decay, and Γ denotes the Gamma function. Individual peak shapes in our model are thus controlled by three parameters. Student-t kernels have shapes that are similar to Voigt profiles, with the degree of freedom corresponding to the mixing weights, and are attractive as (4.5) coincides with the Cauchy distribution when $\nu = 1$, i.e. with a Lorentzian curve in the 1D case, and converges to a Normal distribution as ν approaches infinity. As such they give modelling flexibility to accommodate different peak shapes as well as experimental noise. Since it is difficult to estimate the relative importance of the physical processes leading to the particular strength of Laurentzian and Gaussian in the peak formation via convolution, and since the noise in JRES measurements is not yet well understood, in our applications we fix ν at large value, based on the observation that peaks in the data decay rapidly, and in general the choice of ν should be dictated by the particular experimental conditions.

4.2.2 Modelling of uncatalogued metabolite signal

We model the uncatalogued component of (4.1) using a discrete B-spline wavelet tight frame. Although many applications utilise wavelet bases, redundant wavelet families perform better in rest areas. (Redundancy means that many of the wavelet coefficients are close to zero. Therefore, a high-quality signal approximation can still be achieved even thought those coefficients are disregarded.) Wavelet frames are the easiest to apply among redundant wavelet families. Frames have first been

introduced by Duffin and Schaeffer [40] and gained in popularity since the work of Daubechies et al. [31]. Daubechies constructed a famous orthonormal basis for the space of square-integrable functions, they are named wavelets because of their short fast oscillating waves. While frames are widely used in engineering applications [89, 23], they have been employed less in other fields. (There are many different frames for different applications, e.g. Gabor frames for audio-processing. [6]) For a comprehensive introduction to wavelet frames we refer to Mallat [89] and for further details on the particular systems described in this section to Dong and Shen [37, 38]. Wavelet frames are representation systems consisting of shifts and dilations of compactly supported functions that can provide multiresolution representations of signals, consisting of a low-pass approximation (corresponds to low-pass filters) and high-pass details (corresponds to high-pass filters). They enable localized and adaptive processing of data, e.g. in accordance with prior information, and have successfully been applied in metabolomics. The local support of representation functions makes wavelet expansions a local-influence model, whereas their overlapping support acts as a regularizing mechanism that facilitates stability. Wavelet frames are stable in the sense that small changes in coefficients do not perturb the function significantly and vice versa. Together with the locality and the filtering in low- and high-pass channel information, these characteristics make the expansion coefficients highly interpretable. Beyond stability, localization, and multiresolution, particular wavelet frames offer many advantages in applications. Among the most relevant to this work are the support size of the wavelets, their symmetry (because usually the shape and the form of the signal being filtered matches the general shape of the wavelet and theoretical shape of NMR signal is symmetric) and smoothness properties, as well as the redundancy of the overall system, i.e. its ability to provide sparse and parsimonious representations. Small support size translates to better localization of feature coefficients of the signal and is desirable since it implies lower computational costs and sparse approximation to

local features. Symmetry of the frame elements has the advantage that the corresponding transform can be implemented using mirror boundary conditions without introducing artefacts or increasing the computational burden. This is particularly important in metabolomics applications, since metabolite resonances often appear close to the spectral boundaries. Moreover, metabolomic data has a high amount of inherent local symmetries. To account for the symmetry of peaks in 1D NMR spectra, Astle et al. [3] use Symlet 6 (from the family of Daubechies' least asymmetric wavelets) to model uncatalogued metabolites, as they want to preserve the orthonormality of the representation system. There are several strategies to simultaneously achieve perfect symmetry, small support and smoothness, one of which is to give up orthonormality and to use wavelet tight frames. Tight frames provide stable signal decomposition and reconstruction in the same fashion as orthonormal bases, while having built in redundancy, thus enabling sparser representations than (bi)orthogonal systems and in turn allowing the application of strong shrinkage priors to the transformed coefficients.

Given $\Psi := \{\psi_1, \ldots, \psi_r\} \subset L_2(\mathbb{R})$, a wavelet system can be represented by

$$X(\Psi) := \{\psi_{l,n,k} : 1 \leq l \leq r; n,k \in \mathbb{Z}\},$$

where $\psi_{l,n,k} := D^n T_k \psi_k = 2^{n/2}\psi_l(2^n \cdot -k)$ are *shifts* and *dilations*. If $X(\Psi)$ is *tight frame* for $L_2(\mathbb{R})$, then it is called a *tight wavelet frame* for $L_2(\mathbb{R})$ and the elements of Ψ are called *wavelets*.

Given a separable Hilbert space $\mathscr{H}$ with inner product $\langle \cdot, \cdot \rangle$ and a finite or countable index set I, a sequence $\{g_i\}_{i \in I} \subset \mathscr{H}$ is called a *tight frame* for $\mathscr{H}$ if

$$f = \sum_{i \in I} \langle f, g_i \rangle g_i \quad \text{for all } f \in \mathscr{H}. \tag{4.6}$$

Tight frames thus provide perfect signal reconstruction in the same way as Hilbert space orthonormal bases, without requiring the frame elements to be orthonormal or the coefficients in (4.6) to be unique. (Tight frames are generalizations of the concept of orthonormal basis in Hilbert spaces.) Indeed, the only properties of Hilbert space orthonormal bases that [3] use for their inferential method is (4.6). A tight frame is, in fact, an orthonormal basis if and only if all its elements have unit norm. The coefficients $\{\langle f, g_i\rangle\}_{i\in I} \in \ell_2(I)$ are called the *canonical frame coefficients* of f, where $\ell_2(I)$ denotes the space of square-summable scalar sequences indexed by I. The *analysis operator* of the tight frame maps every signal $f \in \mathscr{H}$ to its sequence of canonical frame coefficients. Its adjoint operator is called the *synthesis operator* and maps $c \in \ell_2(I)$ to the superposition $\sum_{i\in I} c(i)g_i \in \mathscr{H}$. The system $\{g_i\}_{i\in I}$ is a tight frame if and only if the composition of its analysis and synthesis operator is the identity on $\mathscr{H}$.

The elements of a wavelet frame are generated by shifts and dilations of, in general more than one, generators, called *framelets*. (In other words, if a wavelet system $X(\Psi)$ is a frame, its elements are referred as *framelets*. *Framelets* are wave-like functions (wavelets) without an orthonormal basis.) In this article, we use a *discrete B-spline wavelet tight frame*. A spline wavelet is a wavelet constructed by using a spline function, which is a class of functions mainly applied when data interpolation or smoothing is required. This class of frames is widely used in wavelet frame based image restoration and has first been introduced by Ron and Shen [112]. The framelets can be defined via framelet filters, which are the coefficients with which the framelet can be written as a linear combination of shifts of refinement functions. When processing digital images/data, the actual framelet functions are never necessary and only framelet filters are needed. The tight frame is generated via a set of finitely supported *framelet filters* $\{\mathbf{a}^{(l)}\}_{l=1}^{r} \in \ell_2(\mathbb{Z}^d)$ (where here $d \in \{1,2\}$

depending on the dimensionality of our problem) that define a shift-invariant system

$$\{(\mathbf{a}^{(l)}(j-k))_{j\in\mathbb{Z}^d} : l \in \{1,\ldots,r\}, k \in \mathbb{Z}^d\}, \tag{4.7}$$

consisting of all of their integer shifts. Sufficient for the system (4.7) to be a tight frame for $\ell_2(\mathbb{Z}^d)$ is that the filters satisfy the unitary extension principle condition of Ron and Shen [112]. The unitary extension principle of Ron and Shen [112] states that:

Let $\phi \in L_2(\mathbb{R})$ be compactly supported refinable with finite mask h_0 and $\hat{\phi}(0) = 1$. Suppose $\Psi = \{\psi_1,\ldots,\psi_r\}$ are defined by finite masks $h_1,\ldots,h_r$. Then $X(\Psi)$ is a tight frame for $L_2(\mathbb{R})$ provided for all $\xi \in \mathbb{R}$,

$$\sum_{l=0}^{r} |\hat{h}_l(\xi)|^2 = 1 \qquad \text{and} \qquad \sum_{l=0}^{r} \hat{h}_l(\xi)\overline{\hat{h}_l(\xi+\pi)} = 0.$$

If, furthermore, $r = 1$ and $\|\phi\| = 1$ then $X(\Psi)$ is an orthonormal basis of $L_2(\mathbb{R})$.

When the system (4.7) is a tight frame, the analysis and synthesis operators are given via discrete convolutions by

$$\mathbf{W}: \mathbf{u} \in \ell_2(\mathbb{Z}^d) \to \left(\sum_{j\in\mathbb{Z}^d} \mathbf{a}^{(l)}(j-k)\mathbf{u}(j)\right)_{(k,l)\in\mathbb{Z}^d\times\{1,\ldots,r\}} \in \ell_2(\mathbb{Z}^d \times \{1,\ldots,r\}) \tag{4.8}$$

and

$$\mathbf{W}^\top : c \in \ell_2(\mathbb{Z}^d \times \{1,\ldots,r\}) \to \left(\sum_{l=1}^{r}\sum_{j\in\mathbb{Z}^d} c(k-j,l)\mathbf{a}^{(l)}(j)\right)_{k\in\mathbb{Z}^d} \in \ell_2(\mathbb{Z}^d). \tag{4.9}$$

The wavelet systems, corresponding to filters satisfying the unitary extension principle condition via the refinement equations from multiresolution analysis theory, form a wavelet tight frame of functions for $L_2(\mathbb{R}^d)$, for which (4.8) and (4.9) describe the undecimated single level fast wavelet transform. (Undecimated wavelet

transform is a wavelet transform algorithm designed to overcome the lack of translation-invariance of the decimated wavelet transform, which is achieved by removing the downsamplers and upsamplers in the decimated wavelet transform and upsampling the filter coefficients by a factor of 2^{j-1} in the jth level of the algorithm.) Since in our practical application both signals and filters are finite we identify $\ell_2(\mathbb{Z}^d)$ with $\mathbb{R}^{N_C \times N_J}$ and $\ell_2(\mathbb{Z}^d \times \{1,\dots,r\})$ with $\mathbb{R}^{N_C \times N_J \times r}$ for $d = 2$, and with $\mathbb{R}^{N_C}$, respectively $\mathbb{R}^{N_C \times r}$, for $d = 1$. The convolutions in (4.8) and (4.9) are performed using symmetric boundary extensions matching the symmetry of the respective filters. The mask of *B-spline* of order m has the form $\hat{h}_0(\xi) := e^{-ij\frac{\xi}{2}} \cos^m(\xi/2)$, where $j = 0$ for even m and $j = 1$ for odd m. In case $d = 1$, we use the $r = 3$ filters

$$\mathbf{a}^{(1)} = \frac{1}{4}(1,2,1), \quad \mathbf{a}^{(2)} = \frac{\sqrt{2}}{4}(1,0,-1), \quad \mathbf{a}^{(3)} = \frac{1}{4}(-1,2,-1).$$

The lowpass filter $\mathbf{a}^{(1)}$ is the refinement mask of the univariate piecewise linear B-spline $B_2(x) = \max(1-|x|,0)$, since $B_2(x)$ has mask

$$\hat{h}_0(\xi) = (\cos\xi/2)^2 = \frac{1}{4}(e^{i\xi/2} + e^{-i\xi/2})^2 = \frac{1}{4}(e^{i\xi} + 2 + e^{-i\xi}).$$

While the highpass filters $\mathbf{a}^{(2)}$ is a wavelet mask of piecewise linear anti-symmetric ($f(x) = -f(-x)$) framelet and $\mathbf{a}^{(3)}$ is a wavelet mask of piecewise linear symmetric ($f(x) = f(-x)$) framelet. The wavelet masks are

$$\hat{h}_1(\xi) = -i\sqrt{2}(\sin\xi/2)(\cos\xi/2) = \frac{\sqrt{2}}{4}(e^{-i\xi} - e^{i\xi})$$

and

$$\hat{h}_2(\xi) = -(\sin\xi/2)^2 = -\frac{1}{4}(e^{-i\xi} - 2 + e^{i\xi}).$$

In our JRES application, i.e. when $d = 2$, we use the $r = 9$ tensor products of $\mathbf{a}^{(1)}, \mathbf{a}^{(2)}$ and $\mathbf{a}^{(3)}$, i.e., the tight frame we are using consists of the integer-shifts of

nine filters with common support size 3×3.

Note that the number r of filters is dictated by the choice of order for the B-splines and framelets. Our choice of piecewise linear order is motivated by computational tractability. We have experimented with piecewise cubic order, in which case a negligible improvement of performance comes at a computational cost that is unacceptable for applications, since then $r = 5$ for the 1D case and $r = 25$ for 2D case. Moreover, note that we use an undecimated transform, as those perform better than decimated transforms in coefficient processing applications, where shift-invariance of coefficients is desirable due to inaccuracies introduced via positional noise (i.e. noise in multiplet position) and during data acquisition. For details we refer to [89], where the undecimated transform is referred to as the *à-trous* algorithm. Finally, we refrain from using several dilation levels as the consequential increase in data size on the transform side would render the MCMC-algorithm unnecessarily expensive while yielding no significant improvements.

4.2.3 Likelihood

Given measurements $\mathbf{z} \in \mathbb{R}^{N_C \times N_J}$ and targeted metabolites $\mathbf{T}_m := (t_m(x_i, y_j))_{i,j} \in \mathbb{R}^{N_C \times N_J}$ $(m = 1, \ldots, M)$, the likelihood of our model in framelet domain is defined by

$$\mathbf{W}\mathbf{z} = \sum_{m=1}^{M} \beta_m \mathbf{W}\mathbf{T}_m + \boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \varepsilon_{ij\ell} \sim \mathrm{N}(0, \lambda^{-1}), \tag{4.10}$$

where $\boldsymbol{\theta} \in \mathbb{R}^{N_C \times N_J \times r}$ are wavelet frame coefficients of the untargeted signal, r being the number of framelets, and $\boldsymbol{\varepsilon} = (\varepsilon_{ij\ell}) \in \mathbb{R}^{N_C \times N_J \times r}$ are independent identically Normal distributed errors with scalar precision parameter λ. For every $l = 1, \ldots, r$, the matrix $(\theta_{ijl})_{i,j} \in \mathbb{R}^{N_C \times N_J}$ contains the canonical framelet coefficients of the l-th framelet. In the spectral regions specified by the theoretical templates we encounter identifiability issues in the estimation of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)^\top$ as we are attempting

to fit both parametric and nonparametric components. To address this problem we specify localized shrinkage priors. While the identifiability problem in 1D has already been tackled by Astle et al. [3] by imposing a hard thresholding constraint in signal domain, their approach makes computations inefficient and therefore infeasible for the 2D setting. In Sections 4.3 and 4.6, we compare our approach with the prior and wavelet specifications of Astle et al. [3] and highlight the advantages of our method.

4.3 Prior specifications

The problem of identifiability of the regression coefficients $\boldsymbol{\beta}$ of the targeted signal and the frame coefficients $\boldsymbol{\theta}$ of the untargeted signal in the likelihood (4.10) arises because in some regions of the spectrum we attempt to fit both the targeted theoretical templates and the untargeted frame component, while the frame component $\boldsymbol{\theta}$ alone could be used to fit the observed spectra perfectly. Scientific interest is mainly in estimating the relative metabolite concentrations $\boldsymbol{\beta}$. To resolve the unidentifiability problem, therefore, sparse solutions for $\boldsymbol{\theta}$ are preferred, where some of the components of $\boldsymbol{\theta}$ are shrunk towards zero by assigning them a prior distribution with heavy tails and concentration of mass near zero. For 1D NMR spectra, Astle et al. [3] assign a global prior distribution to shrink the wavelet coefficients. Additionally, the authors impose a hard thresholding constraint to components of $W^\top \boldsymbol{\theta}$ (where $W^\top$ denotes the inverse wavelet transform with respect to Symlet 6 wavelets) that fall below a small negative threshold parameter, to which they assign a hyperprior to perform local shrinkage [see Eq. (7) in 3]. The rationale is to prevent the wavelet component of the model to compensate for mismatched metabolites. However, this strategy presents several practical limitations: (i) the components of $\boldsymbol{\theta}$ become highly correlated which significantly slows down convergence of the MCMC algorithm; (ii) the implementation of optimization algorithms, such as gradient-based variational inference, becomes difficult; (iii) the posterior distri-

bution of the wavelet coefficients becomes increasingly complex with growing data size, making it challenging to impose such constraint for JRES spectra which usually are $50-100$ times larger than comparable 1D NMR spectra. For these reasons we opt for an alternative strategy and introduce additional local shrinkage in wavelet frame domain, driven by expert knowledge.

Shrinkage priors: To tackle the unidentifiability problem, we enforce sparse solutions for $\boldsymbol{\theta}$ via global and local shrinkage. There are two main approaches to shrinkage in the Bayesian framework: two component discrete mixture priors (usually with a point mass at zero) known as the spike-and-slab [92, 52] and a variety of continuous shrinkage priors (see, for example,Polson et al. [103], Bhattacharya et al. [10], Piironen et al. [101], Bhadra et al. [9].)The spike-and-slab prior is intuitively appealing as it perform automatic variable selection when the spike is taken to be a delta-spike in the origin and it usually performs well in applications. The main disadvantages of this approach are that the results can be sensitive to prior hyperparameter choices (in particular slab variance and prior on the inclusion probability) and that the posterior inference can be too computationally expensive in high dimensions. On the other hand, continuous shrinkage priors are computationally tractable and offer scalable solutions to complex problems and usually yield similar results to those obtained with a spike and slab approach. Computationally efficient and widely used shrinkage priors are the horseshoe [22], the LASSO [123] and the Student-t prior [124]. We use the horseshoe prior since its flat Cauchy-like tails allow components of $\boldsymbol{\theta}$ to assume large values a posteriori when supported by the data, while its infinitely tall spike at the origin provides strong shrinkage for small entries of $\boldsymbol{\theta}$. We further make use of the localization of the framelets to additionally shrink the framelet coefficients $\boldsymbol{\theta}$ in regions of targeted metabolites.

In more detail, given a global shrinkage parameter τ, the horseshoe prior for θ_{ijl}

can be represented as the scaled mixture of Normals

$$(\theta_{ijl} \mid \mu_{ijl}, \tau) \sim \mathrm{N}(0, \mu_{ijl}^2 \tau^2), \quad \mu_{ijl} \sim \mathrm{C}^+(0, c_{ijl}), \quad \text{for all } i, j, l,$$

where the $\theta_{ijl} \mid \tau$ are conditionally independent and where the local shrinkage parameters μ_{ijl} are assigned half Cauchy distributions. As suggested by Gelman [51], we also assign a half Cauchy distribution to the *global* shrinkage parameter, $\tau \sim \mathrm{C}^+(0, d)$. The hyperparameters c_{ijl} and d govern the amount of local and global shrinkage imposed. For the choice of the c_{ijl} we adopt the following local shrinkage strategy:

(i) Consider spectral regions in the targeted components to which we wish to apply additional local shrinkage in framelet domain, i.e., regions where we want to fit theoretical templates. In these regions, the local shrinkage strategy will shrink the signals from the uncatalogued part, which helps most of the signals being explained by the catalogued part. We suggest that additional local shrinkage should be applied to at least one multiplet of each targeted metabolite. To facilitate accurate posterior concentration estimates, at least one multiplet for each metabolite should deconvolve correctly, and we thus would like to apply extra local shrinkage to multiplets that are less overlapped with strong untargeted signals, so that they can better drive concentration estimation. For instance, in the urine spectrum shown in Figure 4.6 the area around 3.660ppm usually presents severe overlapping, thus, we would not consider extra local shrinkage for multiplets around 3.660ppm. If there is no prior information regarding overlap available, we propose the following two options: (1) For each metabolite, apply extra local shrinkage to the multiplets corresponding to the largest number of protons. The motivation for this strategy is that multiplets with higher number of protons are less likely to be overlapped with stronger signals from untargeted metabolites. For example, the metabolite Valine has four multiplets, located at 0.976ppm, 1.029ppm, 3.601ppm and 2.261ppm. The latter multiplet is not

considered in this work due to its extremely complex structure. The corresponding height ratios of the three remaining multiplets, which are proportional to their number of H-protons, are 3:3:1 and thus we apply extra shrinkage to the two multiplets with the highest number of protons, located at around 1.029ppm and 0.976ppm. (2) Apply extra local shrinkage to all multiplets of the targeted metabolites. This second option is more straightforward and allows robust concentration estimation even when signals of targeted metabolites are partially overlapped with strong signal components of untargeted metabolites. The reason is that the extra shrinkage pushes framelet coefficients towards zero, leaving part of the signal unexplained and leading to an underestimation of the precision parameter λ. For the examples presented in this article we use the first option, as model fitting using this option is often more satisfactory.

(ii) While shrinkage is performed in framelet domain, the spectral regions chosen in the previous step are characterized by parameters $\delta^{\star}_{mu}$ and ζ_{muv} in frequency domain (see Figure 4.1). Using prior information about the uncertainty of these parameters, discussed below, we determine regions, centred around $(\delta^{\star}_{mu}, \zeta_{muv})$, of likely locations for the specified multiplets and identify the index set $\mathscr{I} \times \mathscr{J} \subset N_C \times N_J$ for which (x_i, y_j) belongs to the determined regions. (Recall that the index (i, j) identifies a position in frequency domain.) First, choose low and high shrinkage parameters $0 \leq c_l < c_h$, and let $\omega_{ij} = c_h$ if $(i, j) \in \mathscr{I} \times \mathscr{J}$ and $\omega_{ij} = c_l$ if $(N_C \times N_J) \setminus (\mathscr{I} \times \mathscr{J})$. Next, define the hyperparameters c_{ijl} controlling the local shrinkage of the coefficients θ_{ijl} of the l-th framelet filter ($l = 1, \ldots, r$) located at position $(i, j) \in N_C \times N_J$ via a running average across the filters support with the low and high shrinkage regions described through (ω_{ij}) in signal domain. Specifically, noting that all filters we use have support of size 3×3, consider the index sets $S_{ij} = (\{i-1, i, i+1\} \times \{j-1, j, j+1\}) \cap (N_C \times N_J)$ within the data grid and define

c_{ijl} via

$$\log_{10} c_{ijl} := \frac{1}{|S_{ij}|} \sum_{(m,n)\in S_{ij}} \omega_{mn}.$$

This means that higher shrinkage is applied in the specified regions, with the level of shrinkage weakening towards the boundary of the regions.

Figure 4.2 illustrates the rationale for applying local shrinkage and its effect on the estimation of concentrations in the urine spectrum that we consider in further detail in Section 4.6. We focus on a region in which the targeted metabolites Valine and Isoleucine (templates shown in top panel) are overlapped with an untargeted signal component. The experimentally observed spectrum, shown in black in the the middle and bottom panels, exhibits a multiplet at 0.998ppm that, in theory, could be assigned to either Isoleucine or Valine, a multiplet at 1.045ppm that can only belong to Valine, and signal at around 3.660ppm, part of which could be assigned to Isoleucine. This region is problematic as it is highly overlapped. Note, that there is a multiplet of Isoleucine at around 0.923ppm, but no signal is detected in the given spectrum. Without local (but only global) shrinkage (middle panel), part of the untargeted signal at around 3.660ppm is assigned to Isoleucine, as there the theoretical template for this metabolite presents a mutiplet. In this case, this latter region is driving the estimation of concentration of Isoleucine and the model is relatively insensitive to the information in the region around 0.923ppm. Consequently, the signature template of Isoleucine does not match the shape of the spectral data between 0.920ppm and 1.000ppm. The resulting mismatch between the observed spectrum and the overall targeted metabolite fit is being compensated by a negative frame component such that a perfect overall model fit is achieved even though the Isoleucine concentration is erroneously overestimated. This also leads to coarse overestimation of the concentration for Valine, since the two multiplets at 0.998ppm (overlapping with the multiplet from Isoleucine) and 1.045ppm should have the same intensity. The mutiplet at 1.045ppm is driving the estimation of con-

centration, but it needs to compensate for the fact that signal at 0.998ppm needs to be split between the two metabolites. Altogether, the conflicting information from different parts of the spectrum results in a negative frame component.

Increasing the overall global shrinkage does not resolve this phenomenon, and results in signals in highly overlapped regions getting erroneously over-explained. Moreover, additional global shrinkage would further push the framelet coefficients to zero, leaving relevant parts of the signal unexplained and consequently result in underestimating the precision parameter λ. However, introducing additional local shrinkage to the frame coefficients in regions of targeted metabolites, as described in (i) and (ii) above, can successfully address the problem. As shown in the bottom panel, the region around 0.922ppm is then driving the estimation of concentration of Isoleucine. Because among three regions (0.922 ppm, 0.997 ppm and 3.660 ppm) where signals of Isoleucine are expected, there is least overlapping in the region around 0.922ppm and signals in this region are very weak, which indicates the concentration of Isoleucine should be approximately zero. The region around 1.045ppm is driving the estimation of concentration of Valine. Because, same as that in Isoleucine, among two regions (0.997 ppm and 1.045 ppm), where signals of Valine are expected, there is least overlapping in the region around 1.045ppm and signals in this region are positive, which indicates the concentration of Valine should be some positive value. Due to the extra local shrinkage the frame component captures mainly the untargeted signal and is prevented from compensating for misfitted targeted metabolites.

The remaining prior specifications (for the coefficients of the targeted metabolites and for the precision parameter) are generalizations of the 1D priors used in Astle et al. [3] to our 2D model.

Prior for precision parameter λ: We opt for a conjugate prior and choose a Gamma distribution with shape parameter a and rate $b/2$, where smaller values of a and b correspond to increased uncertainty in the value of λ. For the simulations and

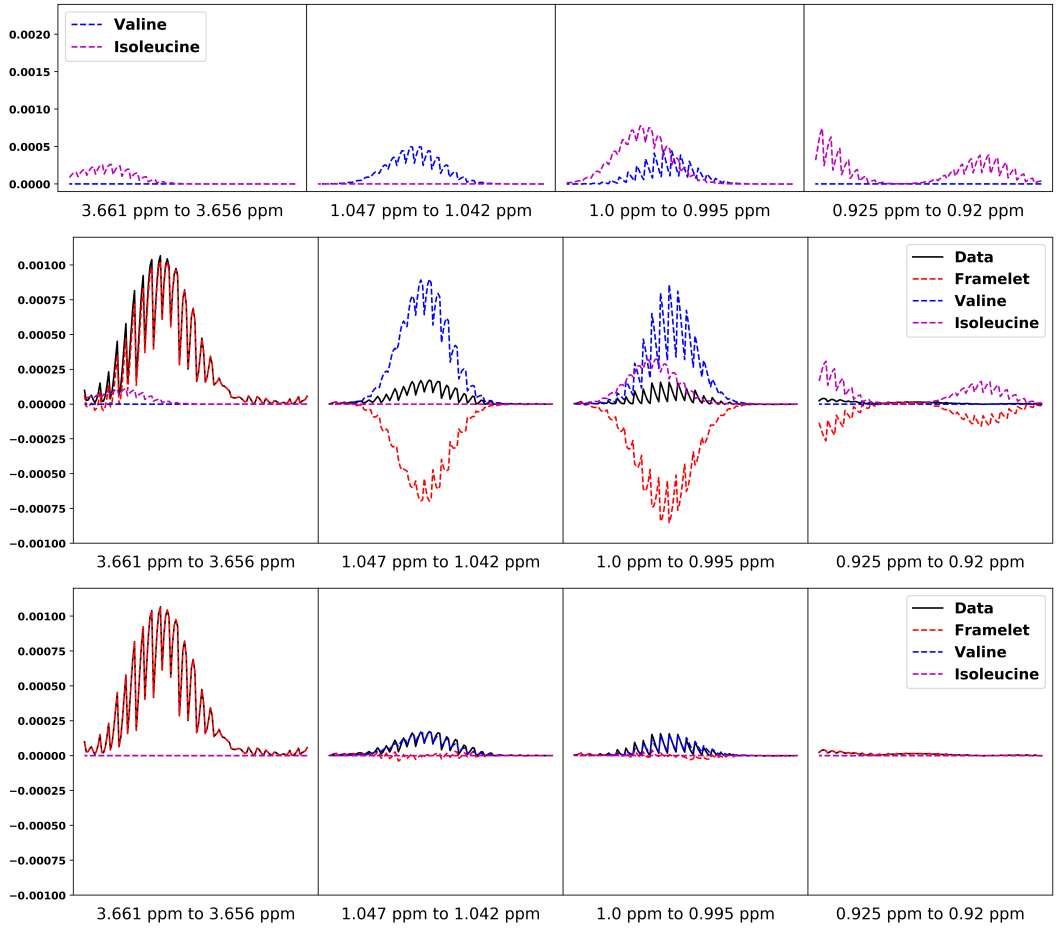

Figure 4.2: Effect of additional local shrinkage applied to framelet coefficients of selected targeted regions. For ease of visualization, spectra are vectorised columnwise and plotted in 2D. On the x-axis we report the chemical shift region of the multiplet, on the y-axis their intensities. The top panel shows the templates of the metabolites Valine and Isoleucine that are targeted. The theoretical template of the multiplet structure of Valine is doublet-doublet-doublet with proton intensity ratio 3:3:1 (recall that we do not include one of the Valine multiplets in the analysis), while that of Isoleucine is triplet-doublet-doublet with proton intensity ratio 3:3:1. Additional local shrinkage is applied in the experiment shown in the bottom panel to the regions of high proton multiplets, i.e. to the first three columns in the lower panel, meaning that estimation is driven by Valine. Compared to the middle panel, in which no additional local shrinkage is applied, this strategy leads to improved accuracy of the concentration estimation for the metabolites.

examples described in this article we choose $a = 10^{-6}$ and $b = 10^{-9}$.

Priors for peak widths: The spectra considered in this article are generated from the biofluids urine and serum. While in this case peak widths change between spec-

tra, their changes are negligible within spectra. We therefore assume that peaks within a spectrum are dependent upon two global peak width parameters σ_1 and σ_2, see (4.5), for which we choose log-Normal distributions with median 1Hz/F and variance 4.6Hz^2/F^2, where F is the operating frequency of the spectrometer in MHz. These priors give good support to a broad region around 1Hz/F, the typical peak widths generated by modern spectrometers [67]. Note that the assumption of common peak widths can easily be relaxed, since local deviations at the metabolite, multiplet or peak level can be modelled via Gaussian random effects on $\log\sigma_1$ and $\log\sigma_2$.

Prior for peak shape: In some applications it might be useful to also assign a prior to the peak shape parameter. Similar to peak widths, peak shapes vary between spectra, but negligibly within spectra. Thus, we assume that peaks within a spectrum depend on a common peak shape parameter ν, see (4.5), to which a log-Normal prior distribution with mean zero and variance 25 can be assigned. This prior gives good support to a broad region around zero. In Section 4.6, we prefer to fix ν.

Priors for multiplets: The parametrization of metabolite signature templates is done in two steps, see (4.3) and (4.4), via linear combinations of multiplets along the chemical shift axis, which in turn arise as linear combinations of Student-t distributions (4.5) along the J-coupling axis. Uncertainty of peak positions can therefore be modelled separately within and between multiplets. The parameters ζ_{muv} and w_{muv}, determining the peak positions on the J-coupling axis and their amplitudes within multiplets in (4.4), can be computed via simple rules from the J-coupling constants J_{mu} (see [67] for details) and may be assumed to be constant across spectra. The multiplet chemical shift parameters $\delta^\star_{mu}$ and J-coupling constants J_{mu} vary slightly between spectra as a result of differing experimental conditions. Empirical estimates $\hat{J}_{mu}$ for J_{mu} and $\hat{\delta}^\star_{mu}$ for $\delta^\star_{mu}$ are published in online databases and can be used to construct an informative prior distribution. The deviations of both J_{mu} and $\delta^\star_{mu}$ from their estimates are local, with smaller variations more likely than

larger ones. Therefore, for each J_{mu} we assign a truncated Normal prior distribution with mean $\hat{J}_{mu}$, variance 7Hz^2, and truncation region $[\frac{1}{2}\hat{J}_{mu}, \frac{3}{2}\hat{J}_{mu}]$. For each $\delta^{\star}_{mu}$ we choose a truncated Normal prior distribution with mean $\hat{\delta}^{\star}_{mu}$, variance 10^{-4}ppm, and truncation region $[\hat{\delta}^{\star}_{mu} - 0.03\text{ppm}, \hat{\delta}^{\star}_{mu} + 0.03\text{ppm}]$. Note that, given specific knowledge about the variability of particular multiplet locations across spectra, it may be appropriate to specify a multiplet- or metabolite-specific alternative for J_{mu} or $\delta^{\star}_{mu}$.

Priors for metabolite abundances: Each coefficient β_m in (4.2) corresponds to the resonance intensity signature of a metabolite and is proportional to the abundance of the metabolite in the biological mixture. Since intensities are positive, the support of the priors for each β_m is restricted to $[0,\infty)$. Conjugacy considerations motivate the use of a truncated Normal prior distribution for each component, i.e. $\beta_m \sim \text{TN}(e_m, 1/s^2_m, 0, \infty)$. This distribution has sufficient flexibility to encode prior information for a wide range of research problems. For the examples presented in this article we choose $e_m = 0$ and $s^2_m = 10^{-6}$ for all $m = 1,\ldots,M$, indicating low prior information.

The details regarding model parameters are shown in Table 4.1.

Observed	Known	To be estimated
$\mathbf{z}$, δ, ζ	w_{muv}, V_{mu}, ρ_{mu}	$\boldsymbol{\beta}$, $\boldsymbol{\theta}$, μ_{ijl}, τ, λ, σ_1, σ_2, ν, J_{mu}, $\delta^{\star}_{mu}$

Table 4.1: Details about Parameters

4.4 MCMC algorithm

We implement an MCMC algorithm to sample from the posterior distribution of the model parameters. Compared to the MCMC strategy in Astle et al. [3], in our setup the MCMC becomes more efficient and easy to implement. For further details on the specific update steps we refer to Supplementary Materials.

We employ Gibbs samplers to update the components of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, both having

truncated Normal conditional distributions, and the precision parameter λ, which has a Gamma distribution. For each of the remaining parameters controlling the targeted and untargeted components of the model we use Metropolis-Hastings updates. Specifically, to update the peak widths parameters σ_1 and σ_2 we use log-Normal proposals. To update the multiplet chemical shift parameter $\delta^{\star}_{mu}$, we propose $\delta^{\star'}_{mu}$ from the truncated Normal distribution

$$\mathrm{TN}\left(\delta^{\star}_{mu}, V^2_{\delta^{\star}_{mu}}, \hat{\delta}^{\star}_{mu} - 0.03\mathrm{ppm}, \hat{\delta}^{\star}_{mu} + 0.03\mathrm{ppm}\right)$$

centred on the current parameter value. Similarly, for the J-coupling constants J_{mu}, we propose J'_{mu} from the truncated Normal distribution

$$\mathrm{TN}\left(J_{mu}, V^2_{J_{mu}}, \frac{1}{2}\hat{J}_{mu}, \frac{3}{2}\hat{J}_{mu}\right).$$

For the local shrinkage parameters μ_{ijl} and the global shrinkage parameter τ we employ Gaussian proposals truncated below at zero. All proposal variances are adapted using the adaptive Metropolis-within-Gibbs algorithm of Roberts and Rosenthal [110], i.e. each proposal variance is tuned to target an acceptance rate of 0.45 by increments and decrements, whose magnitude asymptotically decays at a rate proportional to the inverse of the square root of the iteration number.

Additional Metropolis-Hastings block updates, which prevent the Markov chain from getting trapped in local modes, can be added effortlessly to the described MCMC algorithm. For example, in order to reduce correlation between chains from the targeted and untargeted components of the model in framelet domain, a joint update of a parameter for the targeted component may be introduced. When compared to single parameter updates, such block updates allow the Markov chain to move further, but their acceptance rate is lower. Considering computational efficiency in view of the sizes of JRES spectra, Metropolis-Hastings block updates are therefore

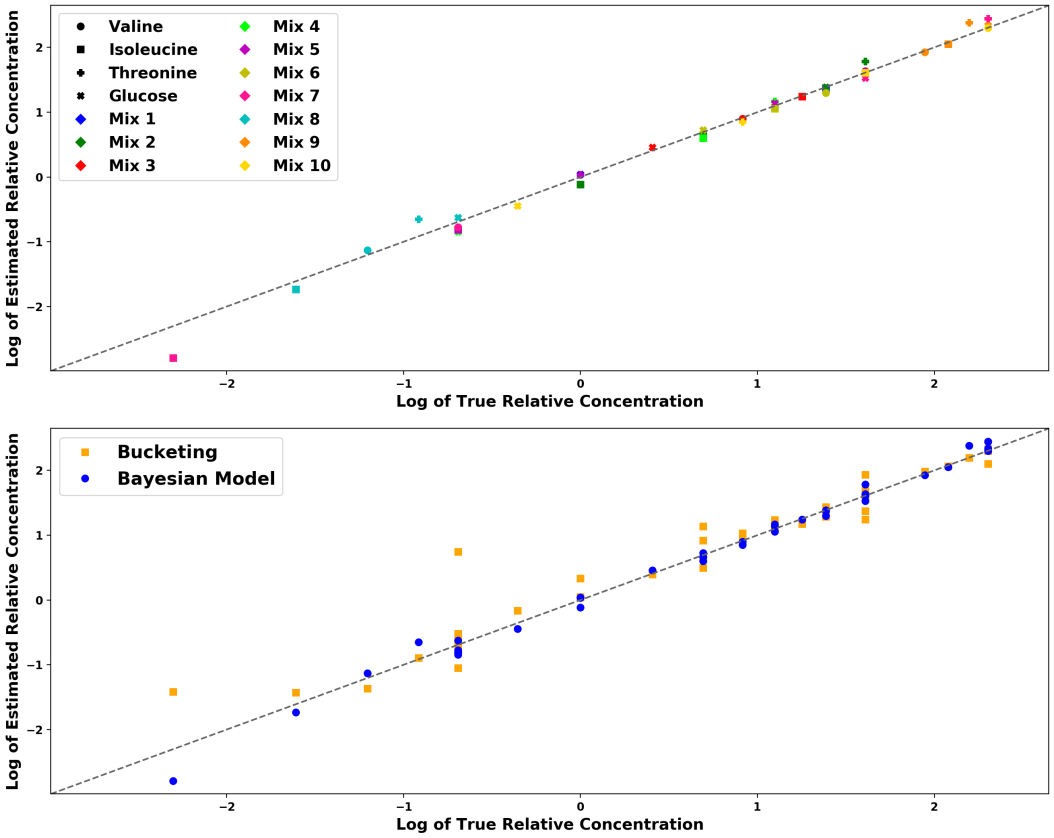

Figure 4.3: Top panel: Comparison between the logarithm of the true relative concentrations and the estimated relative concentrations obtained with our method on the ten mixtures. Bottom panel: Performance comparison between our approach and the bucketing method on the ten simulated biological mixtures.

not utilised in the examples of this article.

4.5 Simulation study

We examine the performance of our method on ten simulated datasets which are created from empirical JRES spectra of the four metabolites Valine (bmse000811), Isoleucine (bmse000884), Threonine (bmse000810) and Glucose (bmse000797) available from the Biological Magnetic Resonance Bank [BMRB, 128]. The synthetic data is generated as follows. First, the empirical spectral template of each metabolite is normalised so that the intensities sum up to one. Then the simulated spectrum is obtained through a linear combination of the four templates with pre-

specified weights. Finally we add Gaussian noise. More specifically, the spectrum of the ith simulated biological mixture is

$$\text{Mix}_i = w_V^i S_V + w_I^i S_I + w_T^i S_T + w_G^i S_G + \boldsymbol{\varepsilon} \quad \text{for } i = 1, \ldots, 10,$$

where w_V^i, w_I^i, w_T^i and w_G^i represent the weights of the Valine, Isoleucine, Threonine and Glucose metabolites, respectively, and S_V, S_I, S_T and S_G represent the respective normalised spectral templates. The weights of the biological mixture can be interpreted as the relative concentrations of each metabolite. Gaussian noise $\boldsymbol{\varepsilon}$ with mean zero and variance 0.001^2 is added to each spectrum. To estimate the relative concentrations of each metabolite in the different mixtures, we also create a baseline spectrum in which all weights are equal to one. We estimate the relative concentration as the ratio between the estimates obtained for the mixture and the ones obtained from the baseline spectrum.

To assess the performance of our model, we compare the logarithm of the estimated relative concentrations with the logarithm of the true relative concentrations. Prior hyperparameters are set as $d = 10^{3.5}$, $c_l = 0$ and $c_h = 5$. The choice of $c_l = 0$ is guided by Carvalho et al. [22], for the choice of d and c_h see Section 4.5 of Supplementary Material. For each dataset, we run $10,000$ iterations of the MCMC algorithm, a burn-in of $5,000$ iterations and thinning every five iterations. Figure 4.3 shows the comparison between true relative concentrations and estimated relative concentrations for the ten biological mixtures. It is evident that our method can estimate the relative concentration very well. Furthermore, we compare our results with those obtained by bucketing (i.e. binning) the spectral data, which is commonly done in metabolic analysis [see, for example, 118]. In this method bins around multiplets corresponding to each metabolite are defined, with bin boundaries validated by an NMR expert. Then relative concentration estimates of each metabolite are obtained by taking the sum of the intensities in the spectral bins

corresponding to each metabolite. From Figure 4.3 it is clear that the bucketing method does not perform as well. Further details on the simulation results and the comparison are presented in Supplementary Material A7.

4.6 Performance on urine and serum spectra

We examine the performance of our method on a urine and a serum dataset. ^{1}H NMR spectra of human urine and serum samples were obtained from healthy participants of the Airwave Health Monitoring Study [43]. The samples were prepared and acquired according to protocols published in Dona et al. [36]. Spectra were acquired at 600MHz with Bruker Ascend configured to the Bruker IVDr specification (Bruker Corporation, Billerica, MA, USA) at 300K (urine) or 310K (serum). 1D NMR spectra were acquired using nuclear Overhauser enhancement spectroscopy (NOESY)-presat using gradients and water suppression (noesygppr1d pulse sequence), a spectral window of 20ppm (urine) or 30ppm (serum), 4s relaxation delay, 10ms mixing time, to a total of 32 transients acquired with 64k data points for urine or 96k data points for serum. 2D JRES data was acquired using the jresgpprqf pulse sequence, with water suppression, a spectral window of 16.6ppm, 2s relaxation delay, 2 scans and 40 increments in the indirect dimension. The spectra were automatically phased and baseline-corrected and chemical shifts were referenced using singlet signal of TSP set at 0ppm (urine) or to the doublet resonance of α-glucose set at 5.23ppm (serum) using Topspin 3.2 software (Bruker Biospin Ltd).

4.6.1 Jres spectra

We demonstrate the performance of our proposed method on the 2D JRES human urine spectrum, with targeted metabolites Valine, Leucine, Isoleucine, Alanine, Lactate and 3-Hydroxy-butyrate. A second performance demonstration on the 2D JRES human serum spectrum is included in Section A9 of the Supplementary Material, and yields results broadly similar to the urine spectrum. A sensitivity analysis

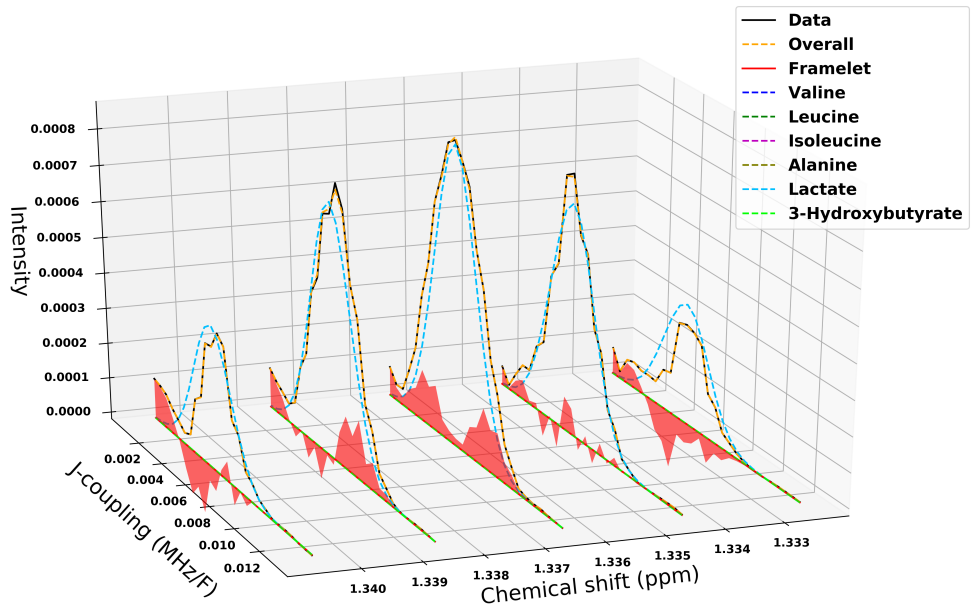

Figure 4.4: Deconvolution surface plot from urine JRES spectrum for the region around 1.337ppm, where the resonance is generated by Lactate. For ease of visualization we plot the fit on a grid of equally spaced points with distance 0.002ppm for the chemical shift axis.

on the 2D JRES human urine spectrum is included in Section A10 of Supplementary Material.

To improve computational efficiency in the quantitative analysis of the test dataset we make use of the theoretical symmetry of 2D JRES spectra with respect to the chemical shift axis and only analyze data with non-negative J-coupling values. Since peaks in the observed spectrum exhibit thin tails, which in some cases drop abruptly to zero due to experimental artefacts, we use bivariate Gaussian distributions, corresponding to bivariate Student-t distributions with large degree of freedom ($\nu = 10,000$). Hyperparameters are set to $d = 10^{3.5}$, $c_h = 5$ and $c_l = 0$. We run the MCMC algorithm for $10,000$ iterations, following upon $5,000$ burn-in iterations, with thinning (selecting every fifth value). The resolution of the urine spectrum is $N_C \times N_J = 436 \times 26$ and the experiment is performed on a laptop with 3.1GHz Intel Core i5 processor, resulting in a run time of 1065 minutes.

Figure 4.8 shows heat maps of the measured spectrum, overall fitting and metabolite

fitting, while Figure 4.4 shows a surface plot of the metabolite fit around 1.337ppm. Along with the additional column-wise 2D plots of the metabolite estimations provided in Figures 3 and 4 in Supplementary Material, they illustrate that our method performs well with respect to goodness of fit, metabolite deconvolution and estimation of relative concentrations. The estimated posterior mean squared error is 7.721×10^{-9}. For Valine (Figure 3, top panel, in Supplementary Material, the first and second multiplets are fitted very well, while the signal from the third multiplet is relatively weak and overlapped with stronger signals from untargeted metabolites, resulting in problematic fitting results. For Leucine (Figure 3, bottom panel, in Supplementary Material, the first and second multiplet should theoretically have the same amplitude (which is not observed); however our method estimates a mid-level concentration of Leucine, resulting in overestimation of the first multiplet and underestimation of the second multiplet. This is reasonable as concentrations are averaged across multiplets. The concentration for Isoleucine (Figure 4, top panel, in Supplementary Material) is close to zero as the signal is very weak at the location of its first multiplet. For Alanine, Lactate and 3-Hydroxybutyrate (Figure 4, bottom panel, in Supplementary Material), the peak shapes differ from Gaussian kernels due to unmodelled experimental conditions. Consequently for each multiplet the high amplitude centre peaks are estimated correctly, while the remaining peaks are slightly underestimated.

As for the convergence of the MCMC, Figures 5 – 10 in Supplementary Material show traceplots of the log-likelihood, of the concentration parameter of Valine and 3-Hydroxybutyrate, of some randomly selected framelet coefficients, of the chemical shift parameter δ, of the J-coupling shift ζ, and of the precision parameter λ. While it can be seen that framelet coefficients and the precision parameter reach convergence quickly, the Markov chain for other parameters, such as the concentration of metabolites, is slow to explore the support of the posterior distribution,

i.e. the Markov chain is mixing slowly. This is to be expected due to overlap and shift of the multiplets. Moreover, it is well known that, when using the horseshoe prior with correlated variables, a main concern is the multimodality of the posterior, which can lead to difficulties in sampling and especially to slow convergence of the MCMC. Nevertheless, from Figure 5 in Supplementary Material it can be seen that the traceplot of the log-likelihood is satisfactory [108].

Finally, in Figures 20 – 25 in Supplementary Material we report the posterior distribution of the concentration parameters and the chemical shift and translation parameters of the six metabolites for the serum and the urine spectra.

4.6.2 Comparison between 1D NMR and 2D JRES deconvolution and quantification, and with bucketing

In the metabolomic literature it is widely accepted that the second dimension provided in 2D JRES spectra can help to mitigate the challenges in the identification and quantification of metabolites in 1D NMR spectroscopy that are mainly due to overlapping ([47, 46]). We illustrate this point by comparing relative concentration estimates using our approach on 1D NMR and on 2D JRES urine spectra from the same sample. Relative concentrations are considered for both datasets since their scaling differs due to data normalization. As baseline metabolite we choose Valine, since it is relatively isolated in both the 1D and the 2D spectra. In Figure 4.5 we compare the estimation results for relative concentrations obtained via our method applied to 1D NMR data, 2D JRES and via the bucketing method. (For the numerical values see Table 2 in Supplementary Material.) Note that bucketing only produces point estimates with no quantification of uncertainty. It is evident that the relative concentration estimates of Leucine, Isoleucine, Alanine and 3-Hydroxybutyrate differ significantly between 1D and 2D spectra. Obviously, 1D NMR leads to much wider 95% credible intervals due to the fact that less information is available in the data. In most cases the credible intervals obtained from 1D and 2D

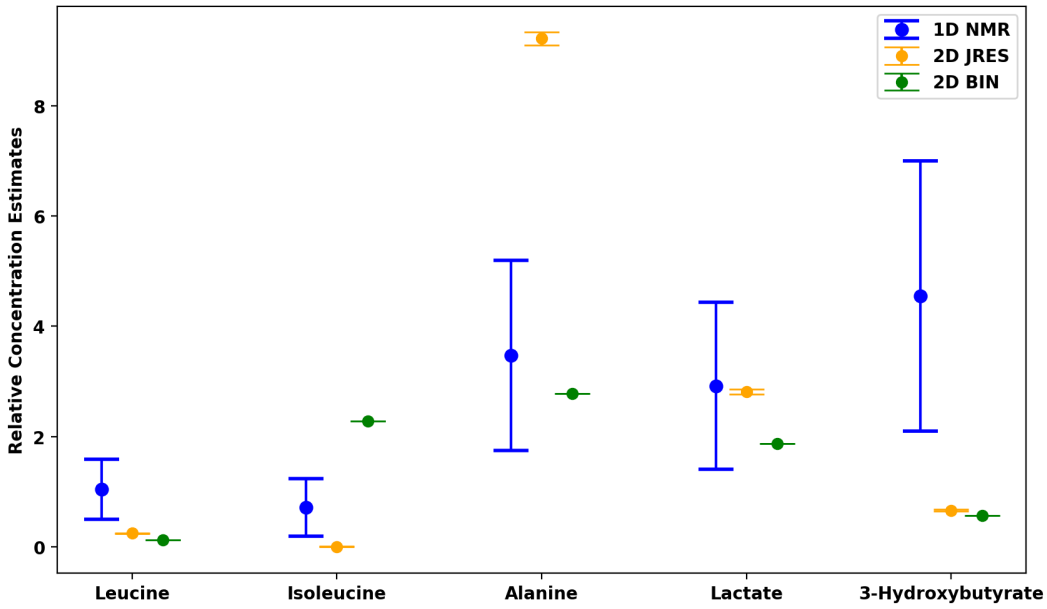

Figure 4.5: Posterior relative concentration estimates and posterior standard deviations using our method on the urine spectrum from 1D NMR measurements, from 2D JRES measurements as well as using the bucketing method on the 2D JRES measurements. Valine is chosen as baseline. For four of the five targeted metabolites the posterior means of the estimates obtained using the second dimension differ by more than 25%. The figure shows 95% credible intervals. Note that bucketing only produces point estimates with no quantification of uncertainty of the estimate.

data do not overlap, clearly showing the potential of 2D NMR spectroscopy. Note that Figure 4.6 shows that the signals from Leucine (around 0.95ppm), Isoleucine (around 0.93ppm, 1.00ppm, 3.65ppm) and 3-Hydroxybutyrate (around 1.20ppm) are severely overlapped with signals from other untargeted or uncatalogued metabolites. This makes identification of signals from targeted metabolites challenging and results in inaccurate estimation of the concentrations. Due to additional information available from the J-coupling dimension, the overlapping issue is less severe in 2D JRES spectra, see Figure 4.8. The underestimation of the concentration for Alanine (around 1.49ppm) from the 1D spectrum stems from fixing J-coupling constants at values slightly different from those observed, as indicated in Figure 4.6 (around 1.49ppm). Moreover, when dealing with urine, a further obstacle to identification

	BATMAN	Our Model
Representation functions for residual spectrum	Symlet 6	Spline framelets
Theoretical peak	Lorentzian	Student-t kernel
Identifiability constraint	Hard constraint through truncation	Horseshoe prior with local shrinkage strategy

Table 4.2: Comparison of modelling strategy between BATMAN and our approach.

and quantification is that some metabolites might be present in the sample at low intensities. In our application the intensities of Valine, Leucine, and Isoleucine signals are lower in urine as compared to their intensities in serum. The signals from Valine can be clearly observed in the JRES spectrum in Figure 4.8, while the signals from Leucine and Isoleucine are present with much lower intensities. This implies that the true concentrations of Leucine and Isoleucine in this urine sample should be much lower than that of Valine. However, the concentration estimates of Valine, Leucine and Isoleucine from the 1D NMR data are close to each other, while the estimates from the 2D JRES data are in line with what would be expected, see Table 2 in Supplementary Material. Traditional bucketing has limitations when being applied to 2D JRES spectra. Firstly, it is difficult to choose the bin boundaries for metabolites in regions of severe overlapping or weak signals, and secondly, severe overlapping can result in overestimation of concentration.

4.6.3 1D NMR spectra and comparison with BATMAN

The R package BATMAN [Bayesian automated analyzer for NMR, see 60, 61] implements the Bayesian method for 1D NMR introduced by Astle et al. [3], but currently cannot be run on 2D NMR data. We therefore compare our method with BATMAN on the 1D human urine data set. Notice that our approach is also suitable to analyse 1D NMR spectra (see Section 4.6.2 below), as it improves on the original strategy adopted in BATMAN. The main modelling differences between our work and the paper by [3] are summarised in Table 4.2. Our improvements have led to a more

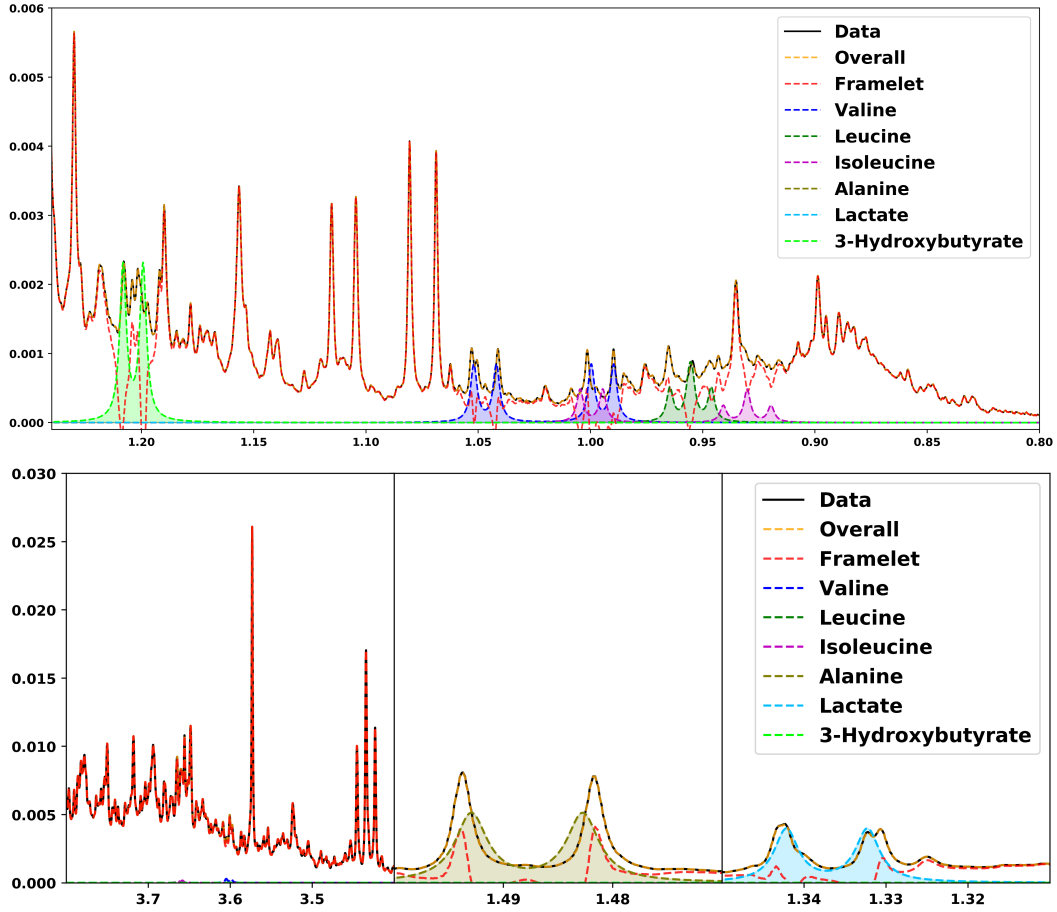

Figure 4.6: Deconvolution of selected regions from the urine 1D NMR data. The x-axis corresponds chemical shift in ppm and y-axis to intensities. The top panel shows resonances generated by Valine, Leucine, Isoleucine and 3-Hydrixybutyrate. The lower middle panel and lower right panel show resonances generated by Alanine and Lactate, respectively. The lower left panel shows resonances generated by untargeted metabolites and weak signals from Valine and Isoleucine.

interpretable model, which is easier to extend to complex set-ups and other 2D NMR techniques and which allows for more efficient computational algorithms.

For a fairer comparison of the efficacy of the untargeted component of our method with BATMAN, we use, like BATMAN does, Lorentzians (i.e. densities of Student-t distributions with one degree of freedom) to model individual peaks, and, when possible, employ the same peak width priors and MCMC strategy as in Astle et al. [3]. Moreover, theoretically, J-coupling constants vary only insignificantly between

spectra, motivating Astle et al. [3] to disregard the fluctuation of J-coupling constants. We therefore also keep J-coupling constants fixed. Parameter values for BATMAN are tuned to yield optimal results for the given data. Specifically, they are set as $a_w = 10^{-9}$, $b_w = 10^{-6}$, $e = 4$, $f = 0.35$, $g = 10^5$, and $h = -0.002$. For our method, we set the shrinkage parameters to $d = 10^{2.5}$, $c_h = 2$ and $c_l = 0$. For both models, $10,000$ iterations of MCMC are performed after $9,000$ burn-in iterations.

Figure 4.6 shows deconvolution of selected region of the urine spectrum obtained with our method. The deconvolution is conditional on the posterior mean of the peak width and chemical shift parameters and is plotted on the same grid as the original spectrum. The original spectral data are shown by the black lines and the framelet component of the model by the red dashed lines. We obtain similar results for BATMAN (results not shown). Indeed, the posterior mean squared error, calculated as the squared difference between the data and the fitted spectrum, is 1.195×10^{-5} for our method and 1.193×10^{-5} for BATMAN, which shows a good performance of both methods. Nevertheless the main limitation of of BATMAN lies in the convergence issues of the MCMC algorithm, due also to the hard constraint that does not allow for an efficient update of the wavelet coefficients. Table 4.3 shows a comparison between the summary statistics of the effective sample sizes (ESS) [106] and of the integrated autocorrelation times (IAC) [28, 73] of the wavelet coefficients for BATMAN and the framelet coefficients for our method. The ESS provides an estimate of the number of independent draws from the posterior distribution of a parameter of interest, while the IAC provides a measure of the efficiency of the sampling algorithm in terms of accuracy of the estimates, with smaller values corresponding to greater efficiency. Using 1000 samples, the mean of the distribution of the ESS of our method is higher than that of BATMAN, indicating a greater number of independent draws in the MCMC for our approach. Since the time requirement of our method is smaller, this implies that the rate of convergence of the untargeted component is faster and the algorithm is more efficient. This is further

		Quantile				Mean	Std dev	Time in secs	Mean /time
		5%	25%	50%	75%				
ESS	BATMAN	90	261	683	906	613	336	7125	0.09
	Our method	98	1000	1000	1000	914	241	5004	0.18
IAC	BATMAN	1.01	1.15	1.45	2.52	2.75	3.95		
	Our method	0.92	0.98	1.04	1.11	2.05	12.24		

Table 4.3: Comparison of effective sample sizes (ESS) and integrated autocorrelation times (IAC) of the coefficients of the uncatalogued signal component between BATMAN and our method. We report summary statistics of the ESS and IAC values of all wavelet/framelet coefficients.

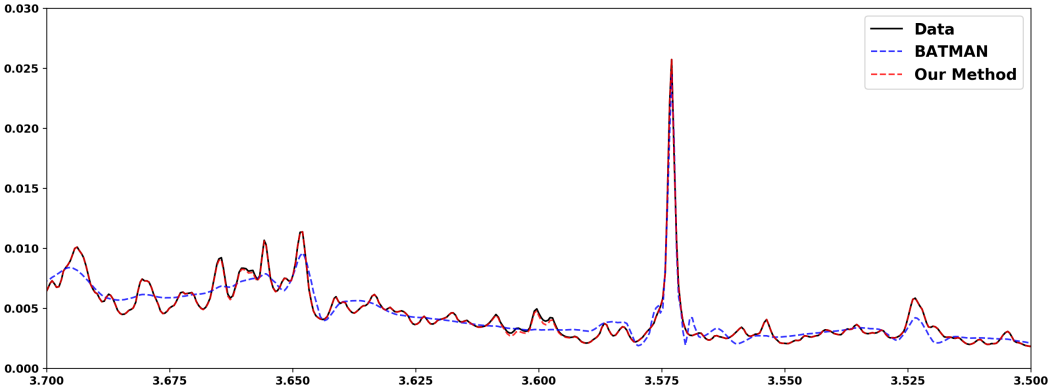

Figure 4.7: Deconvolution of resonances generated by untargeted metabolites for a selected region from a urine 1D NMR spectrum. The x-axis corresponds to chemical shift in ppm and the y-axis to the intensities. The measured spectrum is shown in black, while the B-spline frame component of our model is plotted in red and the Symlet 6 wavelet component of BATMAN in blue.

supported by comparing the IACs: once again, on average, the posterior estimation from our method is more accurate and mixing is improved. Figure 4.7 illustrates that in regions where most of the spectrum is modelled only by framelets, our method improves the fitting compared to BATMAN when using the same number of samples. This is because in the original algorithm in BATMAN the presence of hard constraints included in the model to ensure identifiability lead to lower acceptance rate as they are not always satisfied during MCMC sampling.

4.7 Conclusion

The major advantage of 2D JRES spectra over 1D NMR spectra is that they aid deconvolution, identification and concentration estimation of metabolites by providing information on a second dimension. Presently, there are no automated methods for analyzing 2D JRES spectra that make use of the extensive prior information available in online databases about the physical processes generating the spectral data. Such expert information can be conveniently incorporated into our Bayesian model via specification of informative prior distributions. Analysis of serum and urine spectra, as well as simulations on synthetic data, show that our method can identify resonance peaks correctly. Peak misalignment may occur when a target resonance is overlapped with, or located close to, other strong signals. The latter is inevitable for any method when peaks overlap sufficiently.

A clear advantage of our method is its applicability to JRES spectra of any complex mixture, such as food, soil or petroleum. As prior information on metabolite resonance patterns become more accessible, extensive and precise, a Bayesian method to estimate metabolite concentrations automatically and accurately from 2D JRES spectra has the potential to contribute to many metabolomics research projects. It is, for instance, straightforward to extend our proposed method to a joint model of multiple JRES spectra in which the concentration parameter vector of the targeted metabolites is shared across spectra and treated as a fixed effect, while the remaining parameters in each spectrum are independent. Updates involving components of the concentration vector for the targeted metabolites should then be slightly adjusted from those of the simpler model to reflect the dependence upon multiple spectra. Updates for the remaining parameters remain valid within each spectrum. Moreover, it is in principle straightforward to introduce random effects, with metabolite concentrations varying over spectra, or to incorporate our model into more complex hierarchies in which the main scientific aim might, for instance, be classification or

clustering.

Our method can be used on both 1D and 2D data. The 1D version of our statistical model is more efficient than BATMAN and can be extended to other 2D spectroscopy techniques (e.g. COSY or TOCSY) with the main difference being the type of expert information included in the model. The main limitation of our work is the computational burden of the MCMC algorithm, which limits the applicability of our model to a large collection of spectra. We are developing variational algorithms which can greatly speed up computations, but at the cost of uncertainty evaluation.

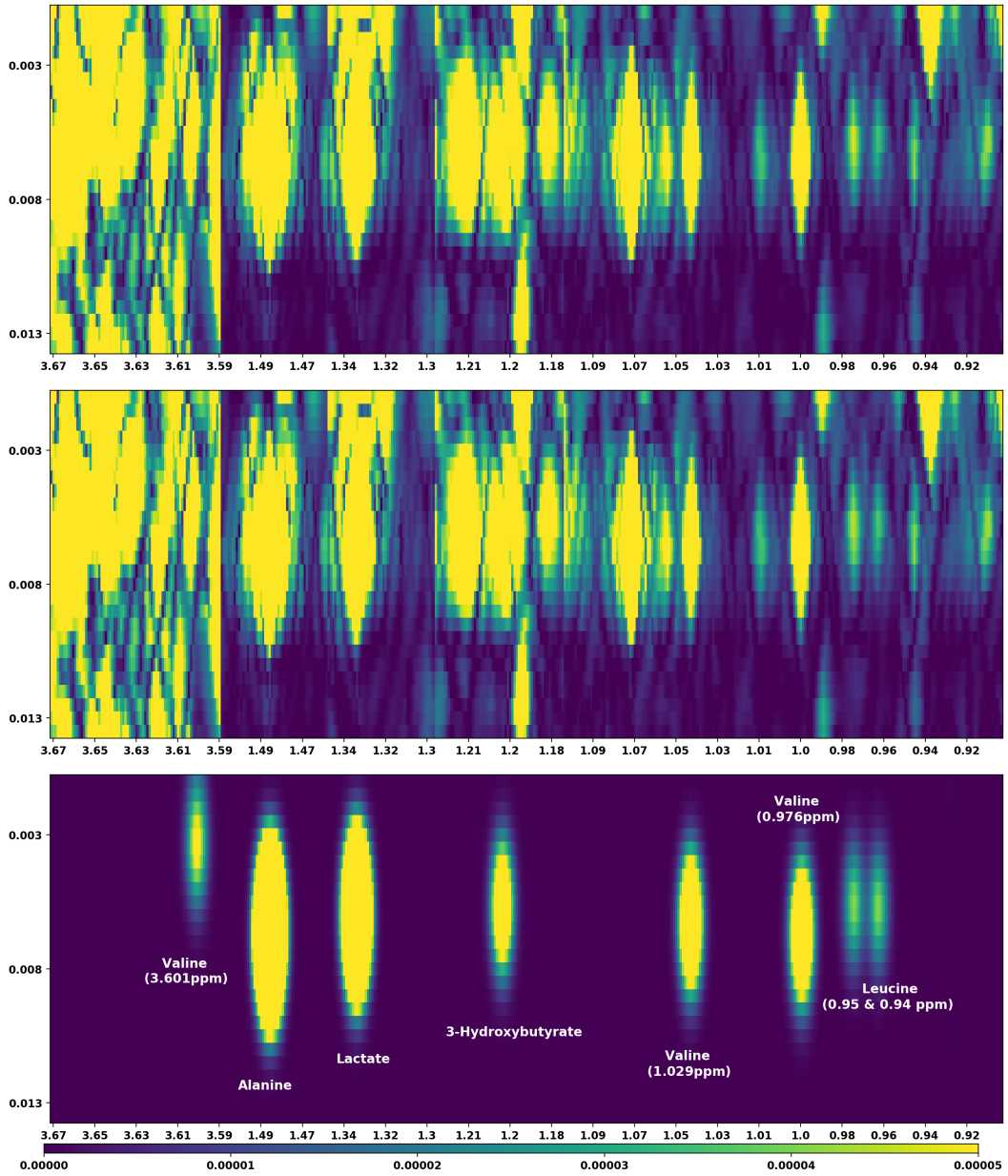

Figure 4.8: Heat maps for intensities from the urine JRES dataset. The x-axis corresponds to chemical shift in ppm, the y-axis to J-coupling in MHz/F. Plots show original data (upper panel), overall fitting, i.e., metabolite and framelet fitting (middle panel), and fitting of metabolites only (lower panel). Multiplets in the lower panel from left to right: Valine (3.601ppm), Alanine, Lactate, 3-Hydroxybutyrate, Valine (1.029ppm), Valine (0.976ppm), Leucine (0.95ppm), Leucine (0.94ppm). The Isoleucine fit is not visible in the lower panel as its concentration estimation is close to zero.

Chapter 5

Final Remarks

5.1 General conclusion

The biggest bottleneck within metabolomics is identification or quantification of metabolites in complex biological mixtures, which is required by almost all experiments in metabolomics. Employing NMR spectroscopy (one of the leading technologies used to capture metabolite data) generates large and heavily structured metabolite dataset. Although analytical approaches and statistical methods for NMR data analysis have been improved, the bottleneck still exists due to a huge diversity of molecular structures and variation of abundance.

Comparing the experimental NMR data with extensively available reference spectra from online databases remains the most reliable approach for metabolite identification and quantification, which have made huge progress because of the development of several databases, e.g BMRB and HMDB. Although these databases provide great value, there are still several challenges to overcome so that we can fully exploit the potential of these databases. One major problem is that the experimental conditions may be inconsistent for different metabolite data contained in these databases. For instance, there might be much variation in pH in different experiments. Besides, online metabolite libraries are continually expanding to include

more metabolite information from an extensive range of biofluids, organisms and experimental conditions. To match unknown resonance with numerous reference standards from these databases, advanced spectral matching algorithms with both accuracy and efficiency are required. Besides, combining and utilising information from different NMR methods, e.g. 1D ^{1}H NMR and JRES, requires close examination.

This thesis tackled three problems involved. In Chapter 2, we design a Bayesian model to effectively estimate the number of protonation sites from sufficient pH titration data for many small molecule metabolites, based on the model of Szakacs et al. [120]. Even when the number of sites was incorrectly estimated, our model is still possible to estimate the chemical shift position of a resonance quite accurately in most cases. The information obtained from the modelling procedure could be valuable for the future development of algorithms for analysis of metabolomic ^{1}H NMR spectra including alignment, annotation and peak fitting. For example, the pH of ^{1}H NMR spectra could be estimated from the positions of a few well known and easily located resonances. This pH information could then be used to predict the chemical shift positions of resonances of other metabolites expected in a sample, which could then help with automated annotation, alignment or peak fitting (as an initial position estimate). The predicted number of protonation sites may also be helpful during the process of identifying unknown compounds, although orthogonal analytical information would almost always be needed in addition.

In Chapter 3, we show that MC-CAVI, as a combination of VI and MCMC, has the potential to improve NMR spectroscopy analysis. Compared with traditionally used MCMC in NMR data analysis, the VI step of MC-CAVI speeds up convergence, makes convergence monitoring easier and provides reliable estimates in a shorter time. Moreover, general Monte Carlo algorithms such as sequential Monte Carlo

and Hamiltonian Monte Carlo can also be incorporated within MC-CAVI effortlessly to further improve the efficiency and accuracy of NMR data analysis. This algorithm can also provides a powerful inferential tool for models particularly in high dimensional settings when full posterior inference is computationally demanding and the application of optimization and of noisy-gradient-based approaches, e.g. BBVI, is hindered by the presence of hard constraints. Besides, MC-CAVI offers a flexible alternative to BBVI. This latter algorithm, although very general and suitable for a large range of complex models, depends crucially on the quality of the approximation to the true target provided by the variational distribution, which in high dimensional setting (in particular with hard constraints) is very difficult to assess.

In Chapter 4, we show that, compared with 1D NMR spectra, the extra information provided by JRES on a second dimension is able to aid deconvolution, identification and concentration estimation of metabolites. Therefore, we design an automated method for analyzing 2D JRES spectra that makes use of the extensive prior information available in online databases about the physical processes generating the spectral data via Bayesian methodology, which has the ability to incorporate expert information conveniently via specification of informative prior distributions. Analysis of serum and urine spectra show that our method can identify resonance peaks correctly. Our method can also be used on 1D NMR data. The 1D version of our statistical model is more efficient than BATMAN and our model can be extended to 2D spectroscopy techniques (e.g. COSY, TOCSY) with the main difference being the type of expert information included in the model.

Metabolomics is a booming field of study and has become a versatile tool, with unceasing technological advancement. Utilising information from difference forms of NMR spectroscopy, development data modelling and incorporating advanced al-

gorithms are crucial to further our comprehension of metabolomics. This thesis contribute to all these three aspects, with the goal to provide assistance to future metabolomics research.

5.2 Future research

In this section, we list some future research directions, which can be beneficial in deepening the understanding of metabolomics and applicability of NMR data analysis.

In future research, one direction is to identify more promising techniques that can be applied in NMR data analysis and combine them to further increase the efficiency of the computational algorithm. One possibility is expectation propagation (EP), which has the potential to reduce computational cost still providing reliable estimate. Introduced by Minka [91], EP soon became a popular approximate Bayesian inference algorithm and was widely applied in statistics, physics [98], deep learning [79] and etc. Suppose we aim to find a tractable distribution, say $q(x)$, to approximate in tractable target distribution, say $p(x)$. In Section 1.5, we discussed variational inference (VI), which identifies an optimal $q^*(x)$ such that

$$q^*(x) = \underset{q(x)\in\mathscr{L}}{\operatorname{argmin}} \operatorname{KL}(q(x) \mid p(x)), \tag{5.1}$$

where $\mathscr{L}$ denotes a family of densities, and KL denotes the Kullback-Leibler divergence. EP, however, achieves the approximation by choosing $q^\star(x)$ such that

$$q^\star(x) = \underset{q(x)\in\mathscr{L}}{\operatorname{argmin}} \operatorname{KL}(p(x) \mid q(x)). \tag{5.2}$$

Swaroop and Turner [119] showed that EP's uncertainty estimates do not collapse pathologically as they do for mean field VI. Therefore, with its superior speed and accuracy, EP has the potential to further improve the computational accuracy and

efficiency of data analysis in metabolomics.

Another possibility is PAC-Bayes, where PAC stands for Probably Approximately Correct. Over the past two decades, PAC-Bayes has been applied as a principled machinery to tackle difficult problems in a wide range of situations such as classification [53], sparse regression [1] and etc. Suppose we have observations $(X,Y) = \{(x_i,y_i)\}_{i=1}^n \in (\mathscr{X},\mathscr{Y})^n$ and we assume that each element (x_i,y_i) $(i = 1\ldots n)$ is randomly sampled from an unknown data generating distribution, say D. In other words, $(X,Y) \sim D^n$. We consider loss functions $l : F \times \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$, where F is a set of predictors $f : \mathscr{X} \to \mathscr{Y}$. Therefore, the empirical risk on the observation (X,Y) is defined as

$$\hat{L}^l_{X,Y}(f) = \frac{1}{n}\sum_{i=1}^{n} l(f,x_i,y_i)$$

and the generalization error over distribution D is defined as

$$L^l_D(f) = \mathbf{E}_{(x,y)\sim D} l(f,x,y).$$

Given the empirical estimate $\mathbf{E}_{f\sim\hat{\rho}}\hat{L}^l_{X,Y}(f)$, PAC-Bayes studies Probably Approximately Correct generalization bounds $\mathbf{E}_{f\sim\hat{\rho}}L^l_D(f)$, where $\hat{\rho}$ is the posterior distribution derived from observations (X,Y). There are two advantages of these generalization bounds: (i) They do not rely on a testing sample; (ii) They are uniformly valid for all $\hat{\rho}$ over F. Therefore, PAC-Bayes could also be useful in data analysis in metabolomics.

Another direction is to extend our Bayesian model in Chapter 4 to other 2D spectroscopy techniques. One possibility is correlation spectroscopy (COSY), which is introduced by Jeener [71] in 1971. Although COSY is the first method in 2D NMR,

it is still one of the most commonly utilised sequences for identifying molecules by their spin-spin couplings and has been applied for metabolite identification in complex mixtures [114, 117], which is usually perceived as difficult. COSY is widely used because it has the capacity to unambiguously identify metabolites of interest in biological mixtures and it does not require long sample pre-treatments.

The second possibility is total correlation spectroscopy (TOCSY), an extension of COSY. Unlike COSY, which only generate correlations between geminal or vicinal protons within a given spin system, TOCSY is able to create correlations between all protons, even when they are distant. Bingol et al. [12] demonstrated the quantification of metabolites via TOCSY based on the assumption that TOCSY transfers can be quantitatively estimated by numerical integration of the Liouville von Neumann equation, which describes the underlying many-spin physics.

Dufour et al. [41] demonstrated COSY's suitability in quantifying pharmaceutical compounds through a simple linear regression model and Bingol et al. [12] quantifies the concentration by simply calculating the numerical integration. Therefore, incorporating experts' knowledge and former experimental results into prior distributions, a Bayesian model for COSY or TOCSY could be an accurate and valuable analytical technique for the quantification of metabolites.

A1 JAGS Code

```
model {
for (i in 1:N){
yobs1[i] ~ dnorm(mu1[i], tau)
yobs2[i] ~ dnorm(mu2[i], tau)
mu1[i] = (deltaa + (incr*deltah1a+decr*deltah1a1)*pow(10,pKa1-pH[i]) +
mod2*(incr*deltah2a+decr*deltah2a1)*pow(10,pKa1+mod2*pKa2-2*pH[i]) +
mod3*(incr*deltah3a+decr*deltah3a1)*
pow(10,pKa1 + mod2*pKa2 + mod3*pKa3 -3*pH[i]))/
(1+pow(10,pKa1-pH[i])+mod2*pow(10,pKa1+mod2*pKa2-2*pH[i]) +
mod3*pow(10,pKa1 + mod2*pKa2 + mod3*pKa3 -3*pH[i]))
mu2[i] = (d2eltaa + (incr*d2eltah1a+decr*d2eltah1a1)*pow(10,p2Ka1-pH[i]) +
mod2*(incr*d2eltah2a+decr*d2eltah2a1)*pow(10,p2Ka1+mod2*p2Ka2-2*pH[i]) +
mod3*(incr*d2eltah3a+decr*d2eltah3a1)*
pow(10,p2Ka1 + mod2*p2Ka2 + mod3*p2Ka3 -3*pH[i]))
/(1+pow(10,p2Ka1-pH[i])+mod2*pow(10,p2Ka1+mod2*p2Ka2-2*pH[i]) +
mod3*pow(10,p2Ka1 + mod2*p2Ka2 + mod3*p2Ka3 -3*pH[i]))
}
mod ~ dcat(p.model[1:3])
inde ~ dcat(p.inde[1:2])
p.model[1] = 1/3
p.model[2] = 1/3
p.model[3] = 1/3
p.inde[1] = 1
p.inde[2] = 0
for (j in 1:3){
xi[j] <-1
}
mod2 <- 1 - (mod == 1)
```

```
incr <- (inde == 1)

decr <- (inde == 2)

mod3 <- (mod == 3)

deltaa ~ dunif(0,10)

deltah1a ~ dunif(deltaa,min(deltaa+1,10))

deltah2a ~ dunif(deltah1a,min(deltah1a+1,10))

deltah3a ~ dunif(deltah2a,min(deltah2a+1,10))

deltah1a1 ~ dunif(max(0,deltaa-1),deltaa)

deltah2a1 ~ dunif(max(0,deltah1a1-1),deltah1a1)

deltah3a1 ~ dunif(max(0,deltah2a1-1),deltah2a1)


d2eltaa ~ dunif(0,10)

d2eltah1a ~ dunif(d2eltaa,min(d2eltaa+1,10))

d2eltah2a ~ dunif(d2eltah1a,min(d2eltah1a+1,10))

d2eltah3a ~ dunif(d2eltah2a,min(d2eltah2a+1,10))

d2eltah1a1 ~ dunif(max(0,d2eltaa-1),d2eltaa)

d2eltah2a1 ~ dunif(max(0,d2eltah1a1-1),d2eltah1a1)

d2eltah3a1 ~ dunif(max(0,d2eltah2a1-1),d2eltah2a1)

pKa1 ~ dunif(1.2,13.7)

pKa2 ~ dunif(1.2,pKa1)

pKa3 ~ dunif(1.2,pKa2)

p2Ka1 ~ dunif(1.2,13.7)

p2Ka2 ~ dunif(1.2,p2Ka1)

p2Ka3 ~ dunif(1.2,p2Ka2)

tau ~ dgamma(10^8,10^4)}
```

A2 Proof of Lemma 1

Proof. Part (i): For a neighborhood of λ^*, we can chose a sub-neighborhood V as described in Assumption 3. For some small $\varepsilon > 0$, the set $V_0 = \{\lambda : \text{ELBO}(q(\lambda)) \geq \text{ELBO}(q(\lambda^*)) - \varepsilon\}$ has a connected component, say V', so that $\lambda^* \in V'$ and $V' \subseteq V$; we can assume that V' is compact. Assumption 3 implies that $M(V') \subseteq V_0$; in fact, since $M(V')$ is connected and contains λ^*, we have $M(V') \subseteq V'$. This completes the proof of part (i) of Definition 1.

Part (ii): Let $\lambda \in V'$. Consider the sequence $\{M^k(\lambda)\}_k$ with a convergent subsequence, $M^{a_k}(\lambda) \to \lambda_1 \in V'$, for increasing integers $\{a_k\}$. Thus, we have that the following holds, $\text{ELBO}(q(M^{a_{k+1}}(\lambda))) \geq \text{ELBO}(q(M(M^{a_k}(\lambda)))) \to \text{ELBO}(q(M(\lambda_1)))$, whereas we also have that $\text{ELBO}(q(M^{a_{k+1}}(\lambda))) \to \text{ELBO}(q(\lambda_1))$. These two last limits give the implication that $\text{ELBO}(q(M(\lambda_1))) = \text{ELBO}(q(\lambda_1))$, so that $\lambda_1 = \lambda^*$. We have shown that any convergent subsequence of $\{M^k(\lambda)\}_k$ has limit λ^*; the compactness of V' gives that also $M^k(\lambda) \to \lambda^*$. This completes the proof of part (ii) of Definition 1. □

A3 Proof of Theorem 1

Proof. Let V_1 be as V' within the proof of Lemma 1. Define $V_2 = \{\lambda \in V_1 : |\lambda - \lambda^*| \geq \varepsilon\}$, for an $\varepsilon > 0$ small enough so that $V_1 \neq \emptyset$. For $\lambda \in V_2$, we have $M(\lambda) \neq \lambda$, thus there are $\nu, \nu_1 > 0$ such that for all $\lambda \in V_2$ and for all λ' with $|\lambda' - M(\lambda)| < \nu$, we obtain that $\mathrm{ELBO}(q(\lambda')) - \mathrm{ELBO}(q(\lambda)) > \nu_1$. Also, due to continuity and compactness, there is $\nu_2 > 0$ such that for all $\lambda \in V_1$ and for all λ' such that $|\lambda' - M(\lambda)| < \nu_2$, we have $\lambda' \in V_1$. Let $R = \sup_{\lambda,\lambda' \in V_1} \{\mathrm{ELBO}(q(\lambda)) - \mathrm{ELBO}(q(\lambda'))\}$ and $k_0 = [R/\nu_1]$ where $[\cdot]$ denotes integer part. Notice that given $\lambda_N^k := \mathscr{M}_N^k(\lambda)$, we have that $\{|\mathscr{M}_N^{k+1} - M(\lambda_N^k)| < \nu_2\} \subseteq \{\lambda_N^{k+1} \in V_1\}$. Consider the event $F_N = \{\lambda_N^k \in V_1 \,;\, k = 0, \ldots, k_0\}$. Under Assumption 4, we have that $\mathrm{Prob}[F_N] \geq p^{k_0}$ for p arbitrarily close to 1. Within F_N, we have that $|\lambda_N^k - \lambda^*| < \varepsilon$ for some $k \leq k_0$, or else $\lambda_N^k \in V_2$ for all $k \leq k_0$, giving that $\mathrm{ELBO}(q(\lambda_N^k)) - \mathrm{ELBO}(q(\lambda)) > \nu_1 \cdot k_0 > R$, which is impossible. □

A4 Gradient Expressions for BBVI

$$\nabla_{\alpha_\vartheta} \log q(\vartheta) = (\vartheta - \alpha_\vartheta) \cdot \exp(-\gamma_\vartheta),$$

$$\nabla_{\gamma_\vartheta} \log q(\vartheta) = -\tfrac{1}{2} + \tfrac{(\vartheta-\alpha_\vartheta)^2}{2} \cdot \exp(-\gamma_\vartheta),$$

$$\nabla_{\alpha_\theta} \log q(\theta) = \big(\gamma_\theta - \tfrac{\Gamma'(\exp(\alpha_\theta))}{\Gamma(\exp(\alpha_\theta))} + \log(\theta)\big) \cdot \exp(\alpha_\theta),$$

$$\nabla_{\gamma_\theta} \log q(\theta) = \exp(\alpha_\theta) - \theta \cdot \exp(\gamma_\theta),$$

$$\nabla_{\alpha_{\kappa_j}} \log q(\kappa_j, \psi_j) = \tfrac{\kappa_j - \alpha_{\kappa_j}}{\exp(2\gamma_{\kappa_j})} + \tfrac{\phi(\frac{\psi_j-\alpha_{\kappa_j}}{\exp(\gamma_{\kappa_j})}) - \phi(\frac{-\psi_j-\alpha_{\kappa_j}}{\exp(\gamma_{\kappa_j})})}{\exp(\gamma_{\kappa_j})(\Phi(\frac{\psi_j-\alpha_{\kappa_j}}{\exp(\gamma_{\kappa_j})}) - \Phi(\frac{-\psi_j-\alpha_{\kappa_j}}{\exp(\gamma_{\kappa_j})}))}, \quad 1 \le j \le n$$

$$\nabla_{\alpha_{\psi_j}} \log q(\kappa_j, \psi_j) = \tfrac{\psi_j - \alpha_{\psi_j}}{\exp(2\gamma_{\psi_j})} + \tfrac{\phi(\frac{2-\alpha_{\psi_j}}{\exp(\gamma_{\psi_j})}) - \phi(\frac{-\alpha_{\psi_j}}{\exp(\gamma_{\psi_j})})}{\exp(\gamma_{\psi_j})(\Phi(\frac{2-\alpha_{\psi_j}}{\exp(\gamma_{\psi_j})}) - \Phi(\frac{-\alpha_{\psi_j}}{\exp(\gamma_{\psi_j})}))}, \quad 1 \le j \le n$$

$$\nabla_{\gamma_{\kappa_j}} \log q(\kappa_j, \psi_j) = \tfrac{(\kappa_j - \alpha_{\kappa_j})^2}{\exp(2\gamma_{\kappa_j})} - 1 + \tfrac{(\psi_j - \alpha_{\kappa_j})\phi(\frac{\psi_j-\alpha_{\kappa_j}}{\exp(\gamma_{\kappa_j})}) + (\psi_j + \alpha_{\kappa_j})\phi(\frac{-\psi_j-\alpha_{\kappa_j}}{\exp(\gamma_{\kappa_j})})}{\exp(\gamma_{\kappa_j})(\Phi(\frac{\psi_j-\alpha_{\kappa_j}}{\exp(\gamma_{\kappa_j})}) - \Phi(\frac{-\psi_j-\alpha_{\kappa_j}}{\exp(\gamma_{\kappa_j})}))}, \quad 1 \le j \le n$$

$$\nabla_{\gamma_{\psi_j}} \log q(\kappa_j, \psi_j) = \tfrac{(\psi_j - \alpha_{\psi_j})^2}{\exp(2\gamma_{\psi_j})} - 1 + \tfrac{(2-\alpha_{\psi_j})\boldsymbol{\phi}(\frac{2-\alpha_{\psi_j}}{\exp(\gamma_{\psi_j})}) + (\alpha_{\psi_j})\boldsymbol{\phi}(\frac{-\alpha_{\psi_j}}{\exp(\gamma_{\psi_j})})}{\exp(\gamma_{\psi_j})(\Phi(\frac{2-\alpha_{\psi_j}}{\exp(\gamma_{\psi_j})}) - \Phi(\frac{-\alpha_{\psi_j}}{\exp(\gamma_{\psi_j})}))}, \quad 1 \le j \le n.$$

A5 MC-CAVI Implementation of BATMAN

In the MC-CAVI implementation of BATMAN, taking both computation efficiency and model structure into consideration, we assume that the variational distribution factorises over four partitions of the parameter vectors, $q(\beta,\delta^*,\gamma)$, $q(\vartheta,\tau)$, $q(\psi)$, $q(\theta)$. This factorization is motivated by the original Metropolis-Hastings block updates in Astle et al. [3]. Let B denote the wavelet basis matrix defined by the transform $\mathscr{W}$, so $\mathscr{W}(B)=\mathbf{I}_{n_1}$. We use v_{-i} to represent vector v without the ith component and analogous notation for matrices (resp., without the ith column).

Set $\mathbb{E}(\theta)=2a/e$, $\mathbb{E}(\vartheta_{j,k}^2)=0$, $\mathbb{E}(\vartheta)=0$, $\mathbb{E}(\tau)=0$, $\mathbb{E}(\mathbf{T}\beta)=\mathbf{y}$, $\mathbb{E}\big((\mathbf{T}\beta)^\top(\mathbf{T}\beta)\big)=\mathbf{y}^\top\mathbf{y}$.

For each iteration:

1. Set $q(\psi_{j,k})=\text{Gamma}\big(c_j+\frac{1}{2},\frac{\mathbb{E}(\theta)\mathbb{E}(\vartheta_{j,k}^2)+d_j}{2}\big)$; calculate $\mathbb{E}(\psi_{j,k})$.

2. Set $q(\theta)=\text{Gamma}(c,c')$, where we have defined,

$$c=a_1+n_1+\tfrac{n}{2},$$

$$c'=\tfrac{1}{2}\Big\{\sum_{j,k}\mathbb{E}(\psi_{j,k})\mathbb{E}(\vartheta_{j,k}^2)+\mathbb{E}\big((\mathscr{W}\mathbf{y}-\mathscr{W}\mathbf{T}\beta-\vartheta)^\top(\mathscr{W}\mathbf{y}-\mathscr{W}\mathbf{T}\beta-\vartheta)\big)+r(\mathbb{E}(\tau)-h\mathbf{1}_n)+e\Big\};$$

   calculate $\mathbb{E}(\theta)$.

3. Use Monte Carlo to draw N samples from $q(\beta,\delta^*_{m,u},\gamma)$, which is derived via (1.14) as,

$$q(\beta,\delta^*,\gamma)\propto\exp\Big\{-\tfrac{\mathbb{E}(\theta)}{2}\big((\mathscr{W}\boldsymbol{T}\beta)^\top\mathscr{W}\boldsymbol{T}\beta-2\mathscr{W}\boldsymbol{T}\beta(\mathscr{W}\mathbf{y}-\mathbb{E}(\vartheta))\big)\Big\}\times p(\beta)p(\delta^*)p(\gamma),$$

where $p(\beta)$, $p(\delta^*)$, $p(\gamma)$ are the prior distributions specified in Section 1.6.1.

- Use a Gibbs sampler update to draw samples from $q(\beta|\delta^*_{m,u},\gamma)$. Draw each component of $\beta=(\beta_m)$ from a univariate Normal, truncated below at zero, with precision and mean parameters given, respectively, by

$$P := s_m + \mathbb{E}(\theta)(\mathscr{W}\boldsymbol{T}_i)^\top(\mathscr{W}\boldsymbol{T}_i),$$

$$(\mathscr{W}\boldsymbol{T}_i)^\top(\mathscr{W}\mathbf{y} - \mathscr{W}\boldsymbol{T}_{-i}\beta_{-i} - \mathbb{E}(\vartheta))\mathbb{E}(\theta)/P.$$

- Use Metropolis–Hastings to update γ. Propose $\log(\gamma') \sim \mathrm{N}(\log(\gamma), V^2_\gamma)$. Perform accept/reject. Adapt V^2_γ to obtain average acceptance rate of approximately 0.45.

- Use Metropolis–Hastings to update $\delta^*_{m,u}$. Propose,

$$(\delta^*_{m,u})' \sim \mathrm{TN}(\delta^*_{m,u}, V^2_{\delta^*_{m,u}}, \hat{\delta}^*_{m,u} - 0.03, \hat{\delta}^*_{m,u} + 0.03).$$

  Perform accept/reject. Adapt $V^2_{\delta^*_{m,u}}$ to target acceptance rate 0.45.

Calculate $\mathbb{E}(\mathbf{T}\beta)$ and $\mathbb{E}\big((\mathbf{T}\beta)^\top(\mathbf{T}\beta)\big)$.

4. Use Monte Carlo to draw N samples from $q(\vartheta,\tau)$, which is derived via (1.14) as,

$$q(\vartheta,\tau) \propto$$
$$\exp\Big\{-\tfrac{\mathbb{E}(\theta)}{2}\Big(\sum_{j,k}\vartheta_{j,k}\big((\psi_{j,k}+1)\,\vartheta_{j,k} - 2\big(\mathscr{W}\mathbf{y} - \mathscr{W}\mathbb{E}(\boldsymbol{T}\beta)\big)_{j,k}\big) + r\sum_{i=1}^{n}(\tau_i - h)^2\Big)\Big\}$$
$$\times\,\mathbb{I}\big\{\mathscr{W}^{-1}\vartheta \geq \tau,\, h\mathbf{1}_n \geq \tau\big\}$$

- Use Gibbs sampler to draw from $q(\vartheta|\tau)$. Draw $\vartheta_{j,k}$ from:

$$\mathrm{TN}\big(\tfrac{1}{1+\mathbb{E}(\psi_{j,k})}\big(\mathscr{W}\mathbf{y}-\mathscr{W}\mathbb{E}(\boldsymbol{T}\beta)\big)_{j,k},\tfrac{1}{\mathbb{E}(\theta)(1+\mathbb{E}(\psi_{j,k}))},L,U\big)$$

  where we have set,

$$L=\max_{i:B_{i\{j,k\}}>0}\frac{\tau_i-B_{i-\{j,k\}}\vartheta_{-\{j,k\}}}{B_{i\{j,k\}}}$$
$$U=\min_{i:B_{i\{j,k\}}<0}\frac{\tau_i-B_{i-\{j,k\}}\vartheta_{-\{j,k\}}}{B_{i\{j,k\}}}$$

  and $B_{i\{j,k\}}$ is the (j,k)th element of the ith column of B.

- Use Gibbs sampler to update τ_i. Draw,

$$\tau_i\sim\mathrm{TN}\big(h,1/(\mathbb{E}(\theta)r),-\infty,\min\big\{h,(\mathscr{W}^{-1}\vartheta)_i\big\}\big).$$

Calculate $\mathbb{E}(\vartheta_{j,k}^2)$, $\mathbb{E}(\vartheta)$, $\mathbb{E}(\tau)$.

A6 Details of MCMC strategy

Denote by $\mathbf{N}(\mathbf{y};\mathbf{b},c)$ the density of a multivariate Normal distribution with mean $\mathbf{b}$ and variance c evaluated at $\mathbf{y}$. Further, denote the combination of targeted metabolites by $\mathbf{T}\boldsymbol{\beta} := \sum_m \beta_m \mathbf{T}_m$. Finally, denote by $\langle\cdot,\cdot\rangle$ the Euclidean inner product on $\mathbb{R}^{N_C \times N_J \times r}$ and by $\|\cdot\|$ its associated norm (i.e. the Frobenius norm). Prior distributions of the parameters are defined in Section 4.3.

Scheme of the MCMC:

1. *Gibbs sampler for* β_m: (Prior location $e_m = 0$ for simplicity). Draw β'_m from univariate Normal distribution, truncated below at zero, with precision parameter $p = s_m + \lambda \|\mathbf{T}_m\|^2$ and mean

$$\frac{\lambda}{p}\left\langle \mathbf{W}\mathbf{T}_m, \mathbf{W}\left(\mathbf{z} - \sum_{k\neq m} \beta_k \mathbf{T}_k\right) - \boldsymbol{\theta} \right\rangle.$$

2. *Gibbs sampler for* θ_{ijl}: Draw θ'_{ijl} from univariate Normal distribution with precision parameter $p = \lambda + (\mu_{ijl}\tau)^{-2}$ and mean

$$(\lambda/p)\,(\mathbf{W}(\mathbf{z} - \mathbf{T}\boldsymbol{\beta}))_{ijl}\,.$$

3. *Metropolis-Hastings update for peak widths*: For $i = 1,2$, propose $\log(\sigma'_i)$ from a $\mathrm{N}(\log(\sigma_i), V^2_{\sigma_i})$. The target distribution is

$$P(\sigma_i \mid \text{rest}) \propto \mathbf{N}(\mathbf{W}\mathbf{z}; \mathbf{W}\mathbf{T}\boldsymbol{\beta} + \boldsymbol{\theta}, 1/\lambda) f(\sigma_i),$$

   where $f(\sigma_i)$ is the prior distribution of σ_i. Perform an accept reject step. Adapt $V^2_{\sigma_i}$ to target acceptance rate 0.45.

4. *Metropolis-Hastings update for multiplet location parameters*: Propose $\delta^{\star\prime}_{mu}$

from truncated Normal distribution

$$\mathrm{TN}\left(\delta^{\star}_{mu}, V^2_{\delta^{\star}_{mu}}; \hat{\delta}^{\star}_{mu} - 0.03\mathrm{ppm}, \hat{\delta}^{\star}_{mu} + 0.03\mathrm{ppm}\right).$$

The target distribution is

$$P(\delta^{\star}_{mu} \mid \mathrm{rest}) \propto \mathbf{N}(\mathbf{Wz}; \mathbf{WT}\boldsymbol{\beta} + \boldsymbol{\theta}, 1/\lambda) f(\delta^{\star}_{mu}),$$

where $f(\delta^{\star}_{mu})$ is the prior distribution of $\delta^{\star}_{mu}$. Perform an accept reject step. Adapt $V^2_{\delta^{\star}_{m,u}}$ to target acceptance rate 0.45.

Propose J'_{mu} from

$$\mathrm{TN}\left(J_{mu}, V^2_{J_{mu}}; \frac{1}{2}\hat{J}_{mu}, \frac{3}{2}\hat{J}_{mu}\right).$$

The target distribution is

$$P(J_{mu} \mid \mathrm{rest}) \propto \mathbf{N}(\mathbf{Wz}; \mathbf{WT}\boldsymbol{\beta} + \boldsymbol{\theta}, 1/\lambda) f(J_{mu}),$$

where $f(J_{mu})$ is the prior distribution of J_{mu}. Perform an accept reject step. Adapt $V^2_{J_{mu}}$ to target acceptance rate 0.45.

5. *Metropolis-Hastings update for shrinkage parameters*: Propose μ'_{ijl} from

$$\mathrm{TN}\left(\mu_{ijl}, V^2_{\mu_{ijl}}; 0, \infty\right).$$

The target distribution is

$$P(\mu_{ijl} \mid \mathrm{rest}) \propto \mathbf{N}(\theta_{ijl}; 0, \mu^2_{ijl}\tau^2) f(\mu_{ijl}),$$

where $f(\mu_{ijl})$ is the prior distribution of μ_{ijl}. Perform an accept reject step.

Adapt $V^2_{\mu_{i,j,l}}$ to target acceptance rate 0.45.

Propose τ' from TN $\left(\tau, V^2_\tau, 0, \infty\right)$. The target distribution is

$$P(\tau \mid \text{rest}) \propto f(\tau) \prod_{i,j,l} \mathbf{N}(\theta_{ijl}; 0, \mu^2_{ijl}\tau^2),$$

where $f(\tau)$ is the prior distribution of τ. Perform an accept reject step. Adapt V^2_τ to target acceptance rate 0.45.

6. *Gibbs sampler for* λ: Draw λ' from Gamma distribution with shape parameter $a + N_C N_J r$ and rate

$$\frac{1}{2}\left(b + \|\mathbf{W}(\mathbf{y} - \mathbf{T}\boldsymbol{\beta}) - \boldsymbol{\theta}\|^2\right).$$

A7 Simulation study details

For the description of the simulation study see Section 4.5. Here, Table 1 reports the true relative concentrations for the ten simulated datasets, their posterior mean estimates obtained with our method, as well as the estimates obtained with the bucketing method. Figure 1 shows the heatmaps of the ten simulated datasets, while Figure 2 shows the baseline spectrum with bin boundaries used for the bucketing method in red.

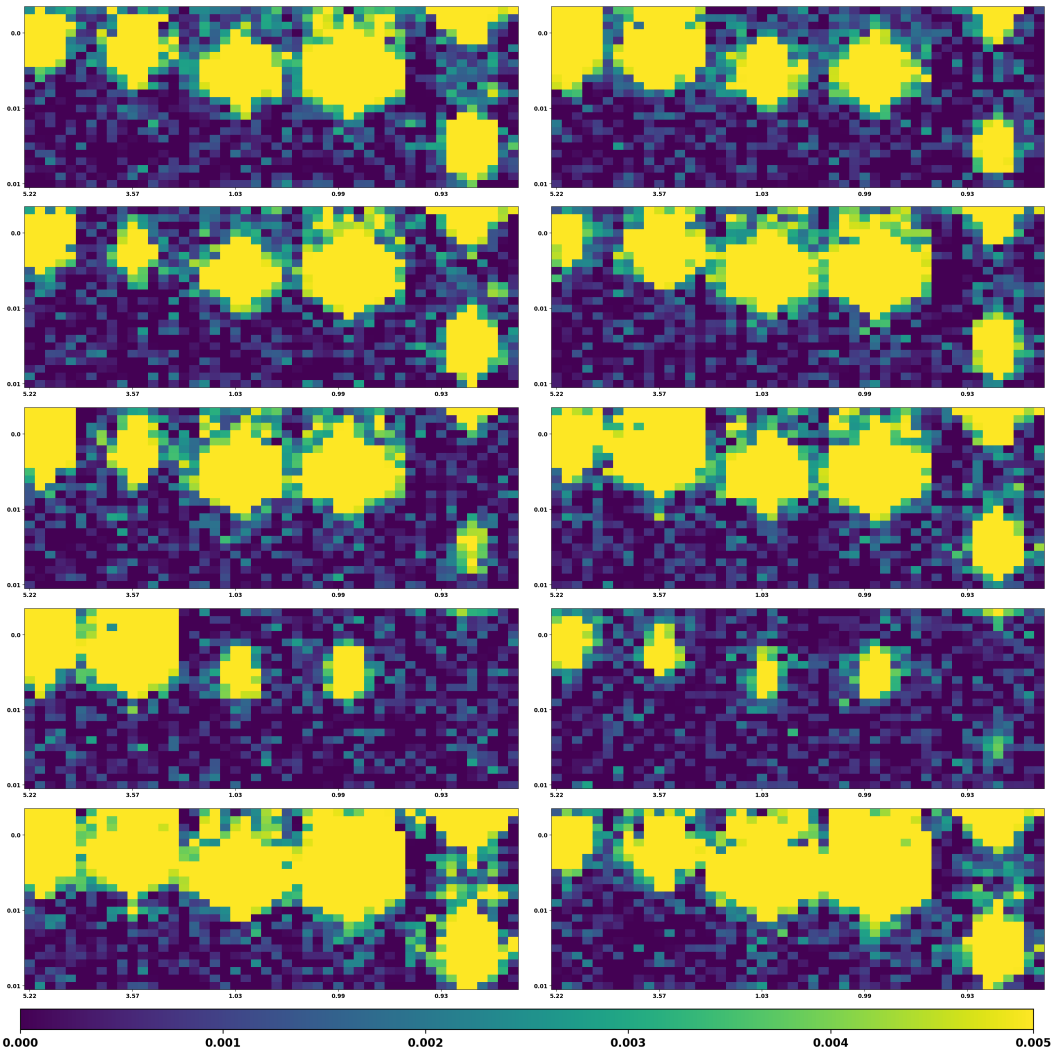

Figure 1: Heatmap of the intensities of the ten simulated biological mixtures (from 1 to 10 row-wise). The x-axis corresponds to chemical shift in ppm, y-axis to J-coupling in MHz/F.

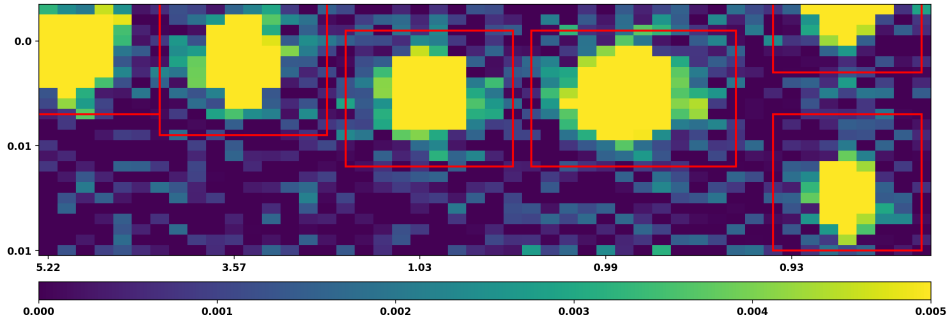

Figure 2: Heatmap of the simulated JRES baseline spectrum with bin boundaries (red). The x-axis corresponds to chemical shift in ppm, y-axis to J-coupling in MHz/F.

Mix		Val	Isol	Thr	Glu	Mix		Val	Isol	Thr	Glu
1	RC	3	4	2	1	6	RC	4	3	10	2
	BAYES	3.0	4.0	2.0	1.0		BAYES	3.7	2.9	11.6	2.1
	BIN	3.3	3.8	2.0	1.1		BIN	3.6	3.5	10.0	2.5
2	RC	2	1	5	4	7	RC	0.5	0.1	10	5
	BAYES	2.0	0.9	6.0	4.0		BAYES	0.5	0.1	11.5	4.6
	BIN	1.6	1.4	5.0	4.2		BIN	0.4	0.2	10.0	5.4
3	RC	2.5	3.5	0.5	1.5	8	RC	0.3	0.2	0.4	0.5
	BAYES	2.5	3.5	0.4	1.6		BAYES	0.3	0.2	0.5	0.5
	BIN	2.8	3.2	0.5	1.5		BIN	0.3	0.2	0.4	0.5
4	RC	5	2	3	0.5	9	RC	7	8	9	10
	BAYES	5.0	1.8	3.2	0.4		BAYES	6.9	7.8	10.9	10.4
	BIN	4.0	3.1	3.0	0.6		BIN	7.3	7.9	9.0	10.3
5	RC	5	0.5	1	3	10	RC	10	5	2.5	0.7
	BAYES	5.1	0.4	1.0	3.1		BAYES	10.0	5.0	2.3	0.6
	BIN	3.5	2.1	1.0	3.0		BIN	8.2	7.0	2.5	0.8

Table 1: True relative concentrations (RC), posterior estimates of relative concentrations obtained with our model (BAYES) and estimates obtained by bucketing/binning (BIN) for ten simulated biological mixtures of Valine (Val), Isoleucine (Iso), Threonine (Thr) and Glucose (Glu).

A8 Additional figures and convergence from urine data

Here we show two additional deconvolution results (Figures 3 and 4) and display traceplots (Figures 5 – 10) for assessing the convergence of the MCMC algorithm when run on the urine dataset of Section 4.6.1 of the main manuscript. Moreover, Table 2 reports the numerical values of the estimates corresponding to Figure 4.5 obtained applying our method on 1D and 2D mesurements, and bucketing the 2D measurements of urine dataset of Section 4.6.1 of the main manuscript.

		Valine	Leucine	Isoleucine	Alanine	Lactate	3-Hydroxy-butyrate
1D NMR	RC	1.000	1.045	0.712	3.468	2.918	4.550
	SD		0.274	0.262	0.862	0.757	1.228
2D JRES	RC	1.000	0.246	0.003	9.220	2.813	0.659
	SD		0.003	0.004	0.061	0.025	0.008
2D BIN	RC	1.000	0.120	2.279	2.776	1.867	0.563

Table 2: Posterior relative concentration estimates (RC) and posterior standard deviations (SD) using our method on urine spectra from 1D NMR measurements as compared to 2D JRES measurements and to 2D bucketing/binning. Note that no standard deviation is available for the bucketing/binning method. Valine is chosen as baseline. For four of the five targeted metabolites the posterior means of the estimates obtained using the second dimension differ by more than 25%**.**

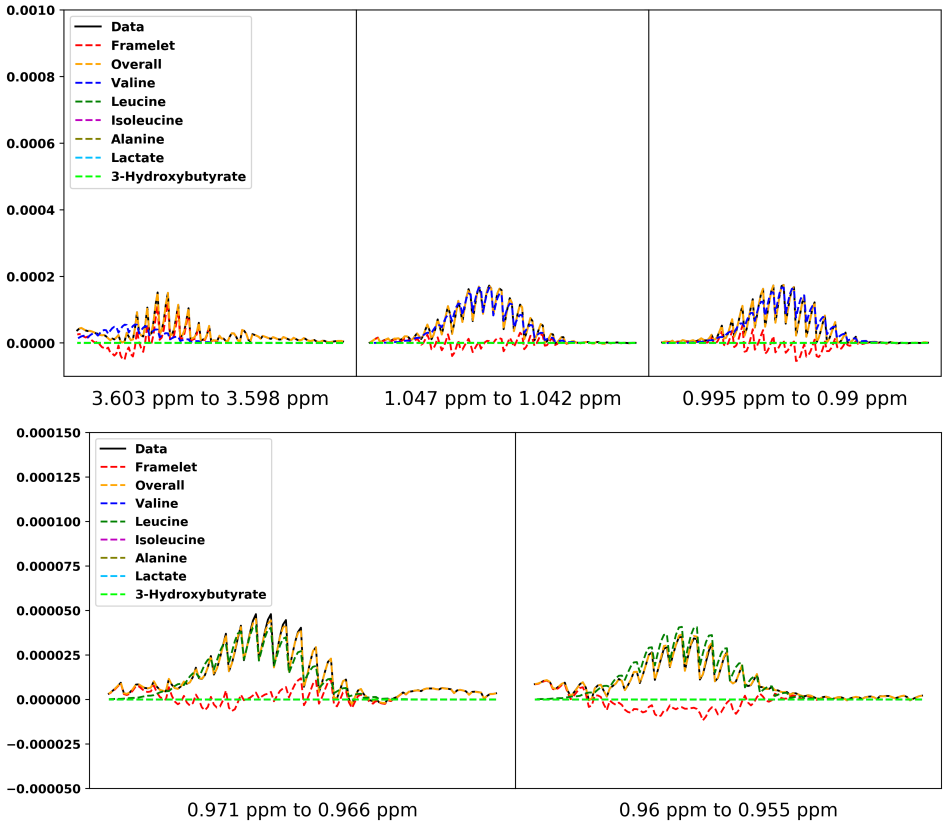

Figure 3: Deconvolution of selected regions from the urine JRES data. Panels show resonances generated by Valine (top panel) and Leucine (bottom panel). On the x-axis we report the chemical region of the multiplet. On the y-axis we report the intensity of the multiplet. The data is vectorised columnwise and plotted in 2D. Original data is displayed in black, untargeted component of the model is displayed in red.

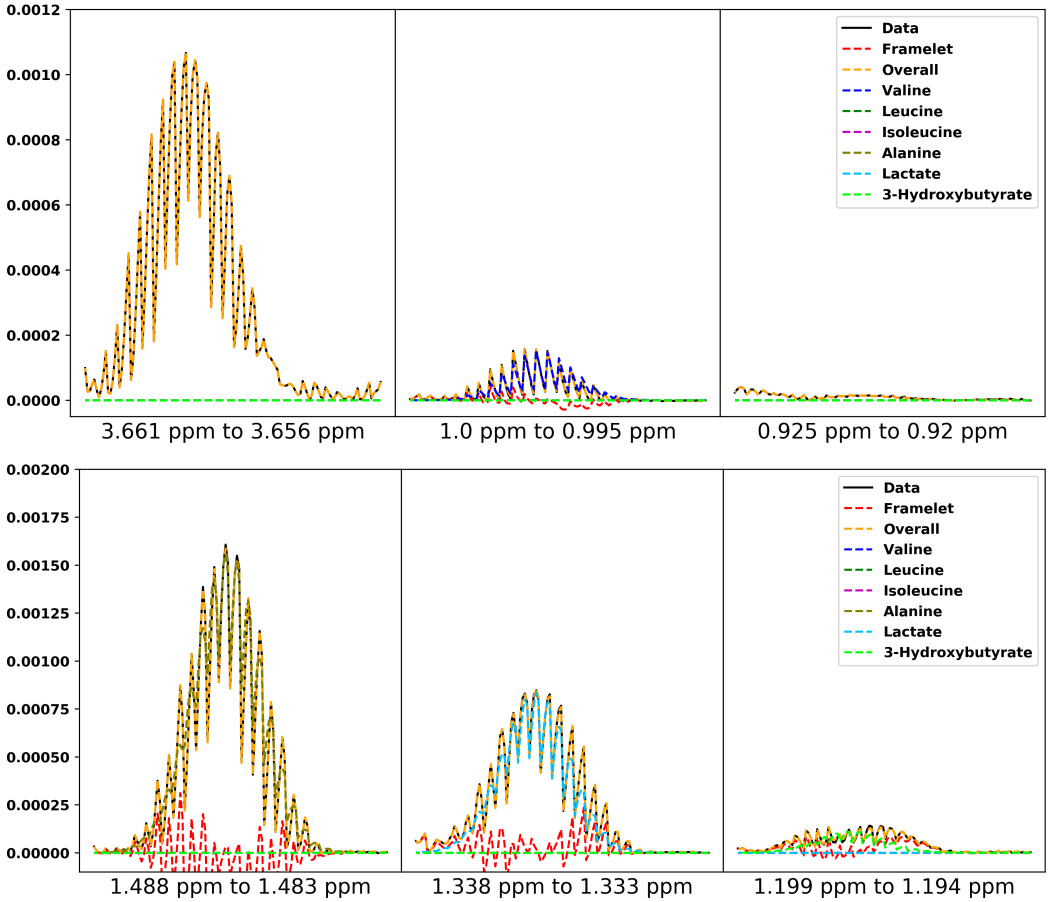

Figure 4: Deconvolution of selected regions from the urine JRES spectrum. Top panel shows resonances generated by Valine and Isoleucine. The latter is not present at a detectable level. Bottom panel shows resonances generated by Alanine, Lactate and 3-Hydroxybutyrate. On the x-axis we report the chemical shift region of the multiplet. On the y-axis we report the intensity of the multiplet. The data is vectorised columnwise and plotted in 2D.

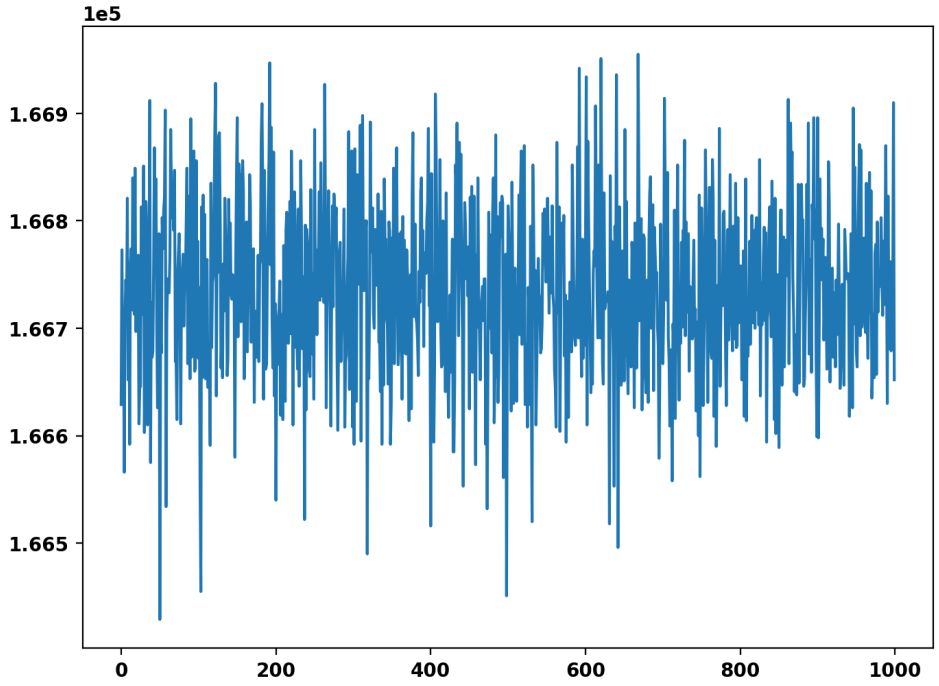

Figure 5: Traceplot of the log-likelihood.

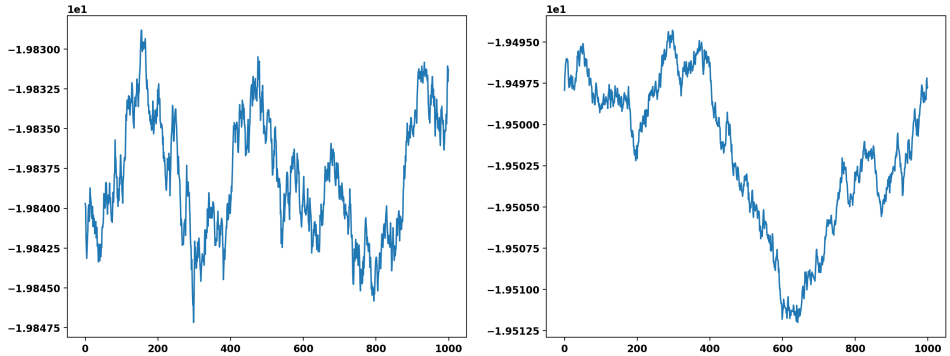

Figure 6: Traceplot of $\log\beta$ (concentration parameter) of Valine (right panel) and 3-Hydrixybutyrate (left panel). The x-axis corresponds to the number of iterations, the y-axis corresponds to the logarithm of the sample value.

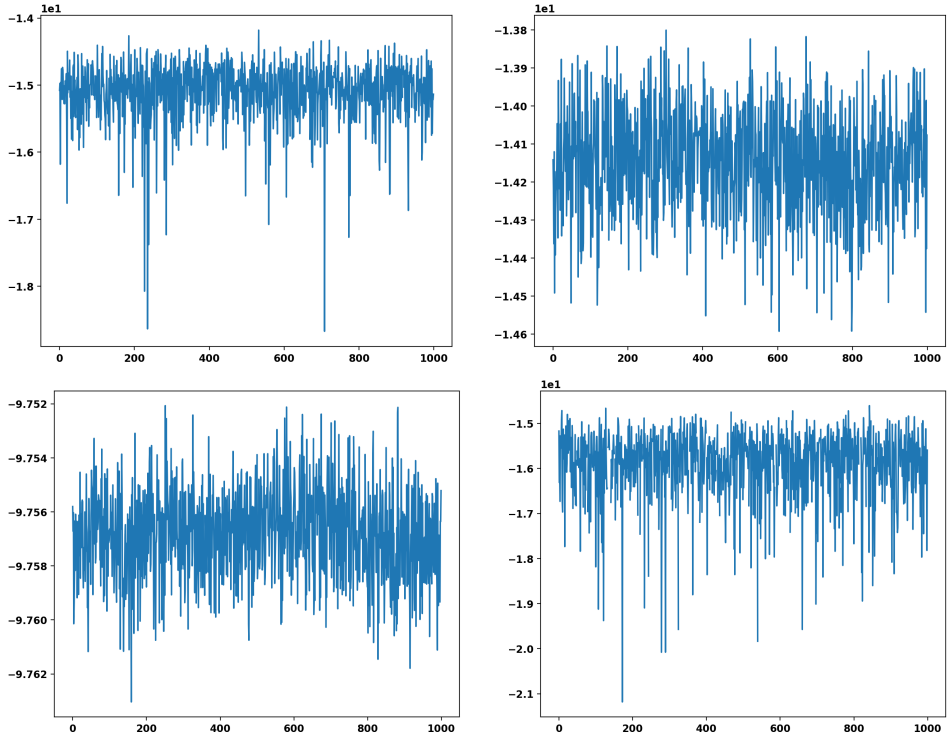

Figure 7: Traceplots of $\log \theta_{ijl}$ for four randomly chosen framelet parameters. The x-axis corresponds to the number of iterations, the y-axis corresponds to the logarithm of sample values.

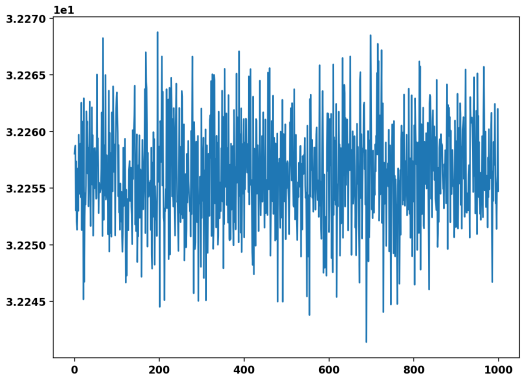

Figure 8: Traceplot of $\log \lambda$ (precision parameter). The x-axis corresponds to the number of iterations, the y-axis corresponds to the logarithm of sample value

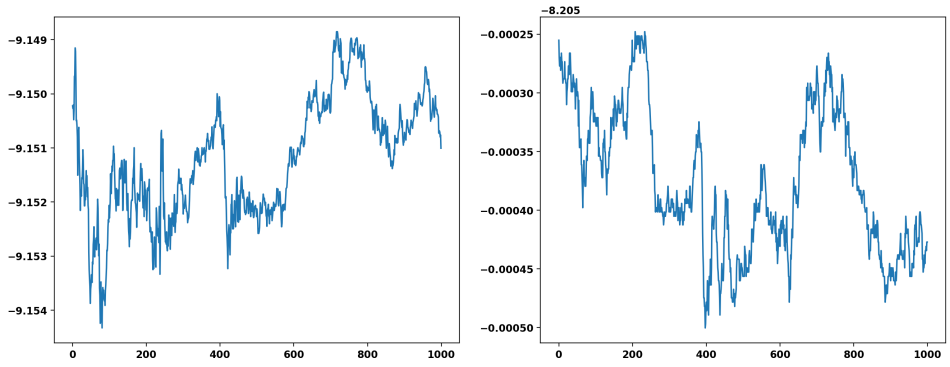

Figure 9: Traceplot of $\log \delta$ **(chemical shift parameter) of Valine** 1.029**ppm (left panel) and Leucine** 0.95**ppm (right panel). The** x**-axis corresponds to the number of iterations, the** y**-axis corresponds to the logarithm of sample values.**

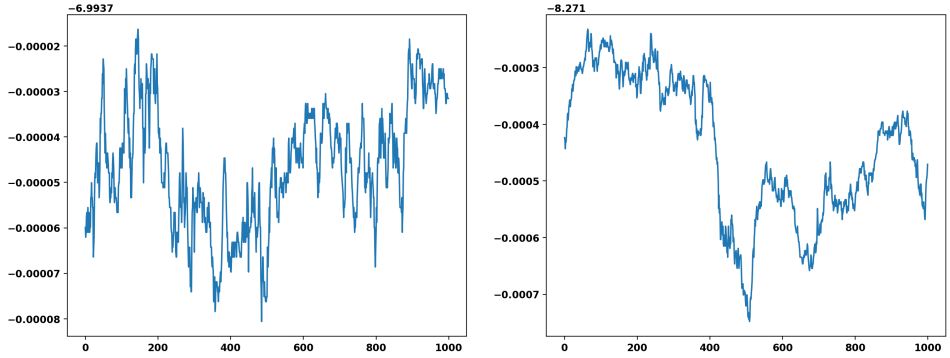

Figure 10: Traceplot of $\log \zeta$ **(**J**-coupling parameter) of Isoleucine** 3.65**ppm (left panel) and Alanine (right panel). The** x**-axis corresponds to the number of iterations, the** y**-axis corresponds to the logarithm of sample values.**

A9 Serum metabolite quantification from JRES spectrum

We demonstrate the performance of our method on a serum JRES spectrum of resolution $N_C \times N_J = 160 \times 26$, with targeted metabolites Valine, Leucine and Isoleucine. We run the MCMC algorithm for $10,000$ iterations, following upon $5,000$ burn-in iterations, with thinning (selecting every fifth value). The experiment is performed on a laptop with 3.1GHz Intel Core i5 processor, resulting in a runtime of 44 minutes. The posterior estimate of the mean squared error is 1.711×10^{-8}. Figure 12 shows heat maps of the spectrum, overall model fitting and metabolite fitting. Our method correctly identifies the the targeted metabolites. Figure 11 shows a surface plot in the region around 0.958ppm, while Figure 13 shows a vectorized 2D plot of our fitting. Figures 11 and 13 show that the overall quantification result is especially accurate for Leucine. For Valine, theoretically, the amplitude ratios of the three mutiplets should be 3:3:1. Consequently, the first and second multiplet are underestimated since the height of the third multiplet is constrained by the data. For Leucine, the measured spectral peaks differ from Gaussian kernels due to unmodelled experimental conditions. Therefore, for both multiplets, high amplitude center peaks are slightly overestimated while the remaining peaks are slightly underestimated. In the case of Isoleucine, the first multiplet is estimated correctly, while the second multipet is underestimated and the third multiplet is overestimated slightly for similar reasons.

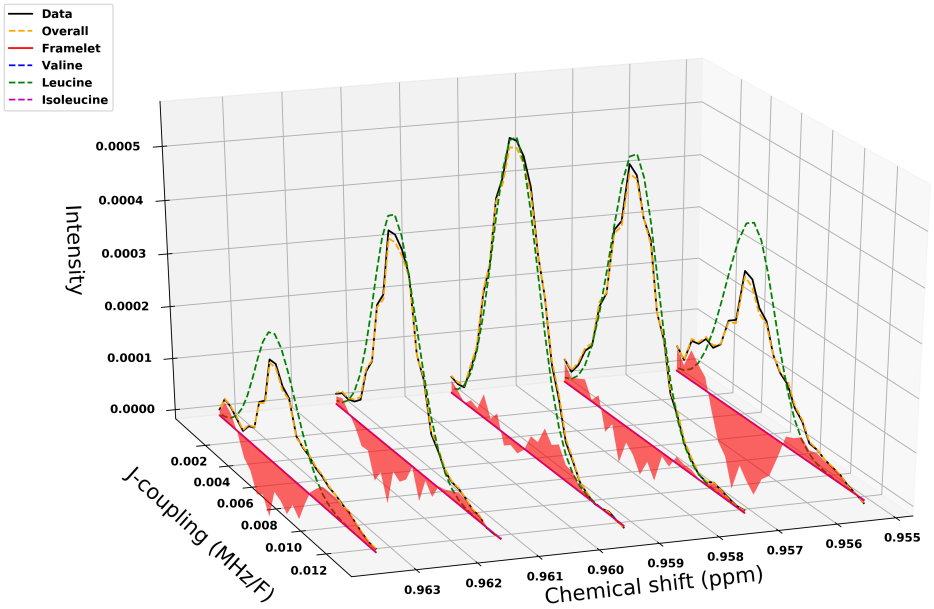

Figure 11: Surface plot of deconvolution for region around 0.958ppm from the serum JRES spectrum. In this region the resonance is generated by the second multiplet of Leucine. For ease of visualization we plot the fit on a ppm-grid of 0.002 equally spaced points.

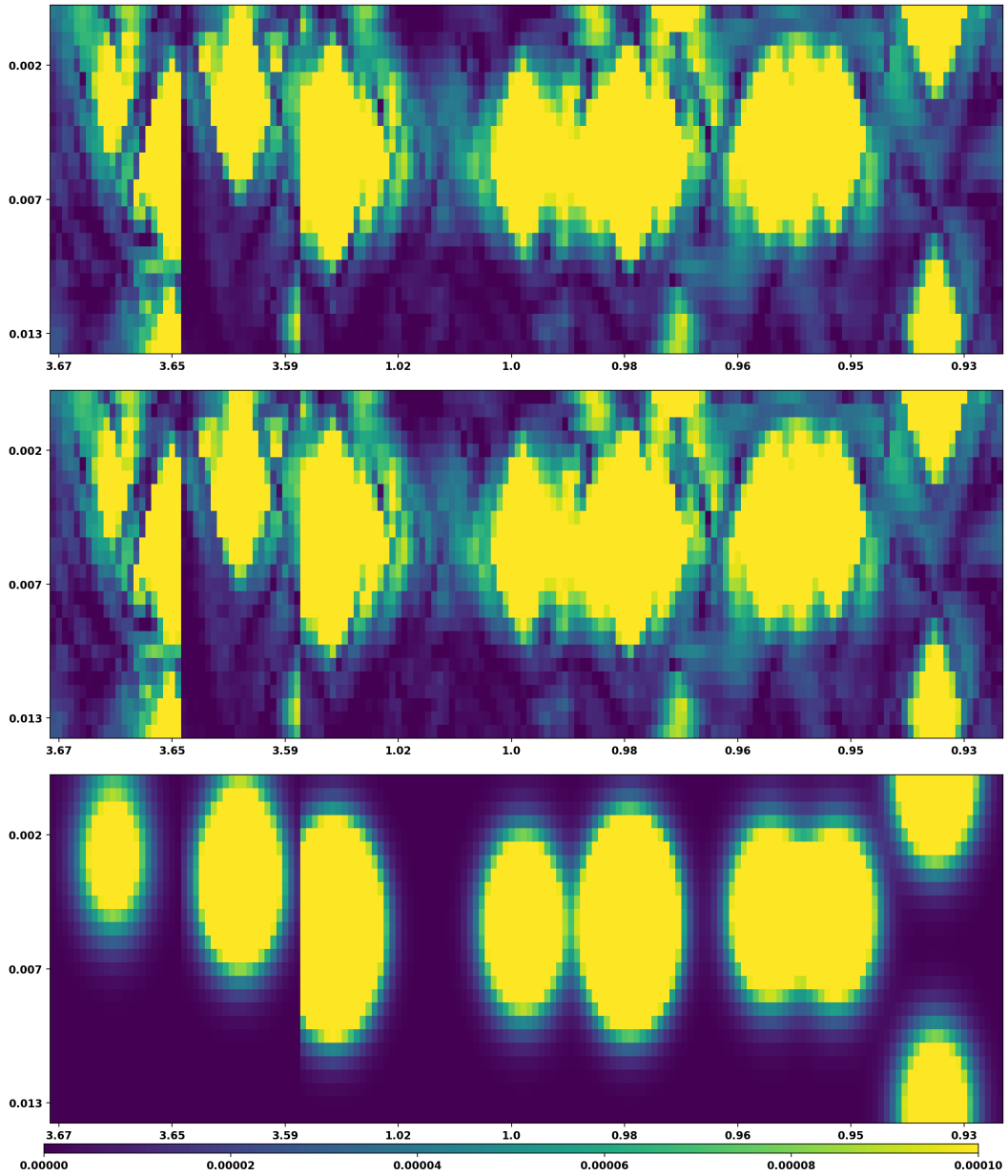

Figure 12: Heat maps for intensities from serum JRES spectrum. The x-axis corresponds to chemical shift in ppm, y-axis to J-coupling in MHz/F. Upper panel shows the original data, middle panel shows the overall fitting (metabolite fitting and wavelet fitting), and lower panel shows metabolite fitting only. The multiplets in the lower panel from left to right are: Isoleucine (3.66ppm), Valine (3.60ppm), Valine (1.03ppm), Isoleucine (1.00ppm), Valine (0.98ppm), Leucine (0.95ppm), Leucine (0.94ppm), Isoleucine (0.93ppm).

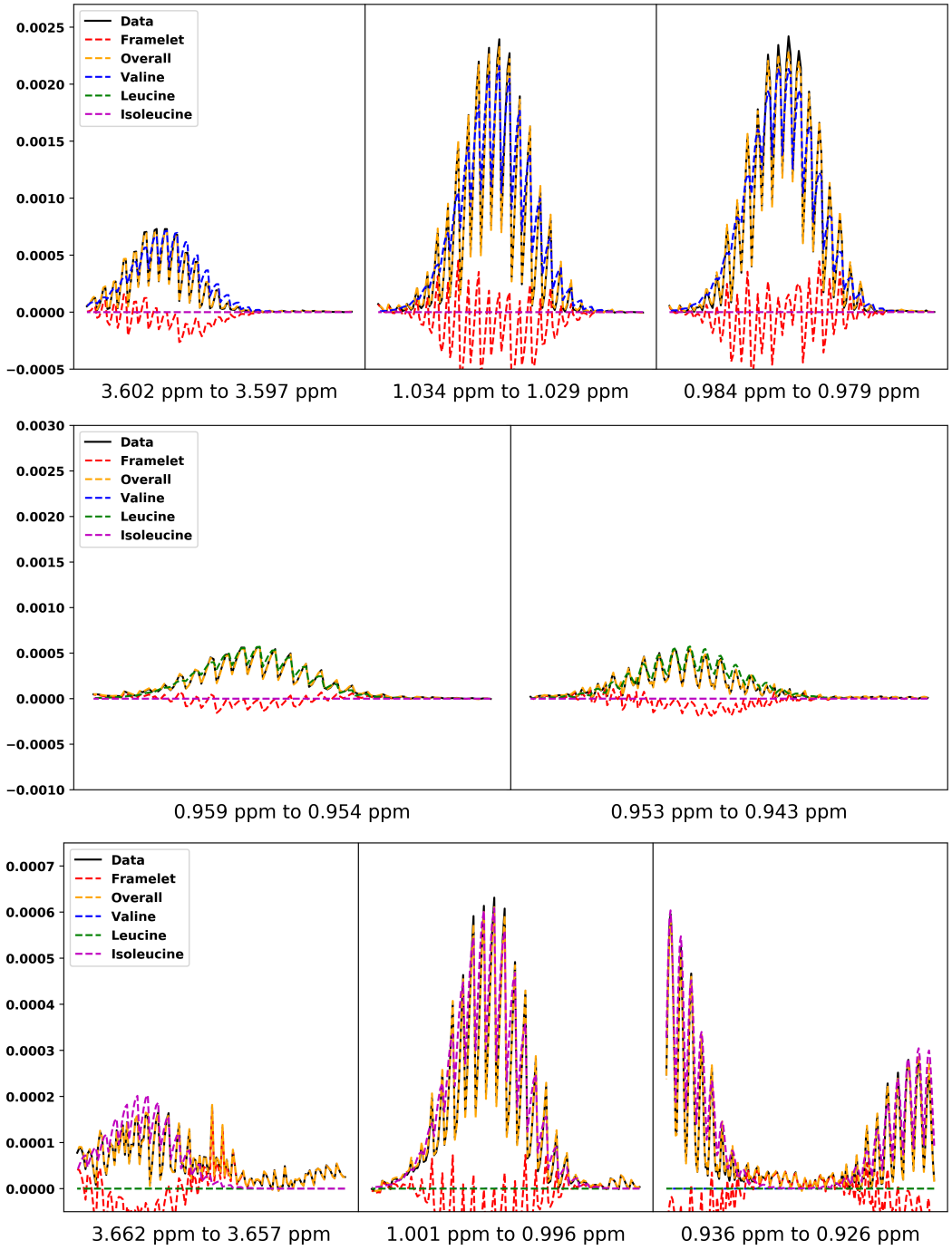

Figure 13: Deconvolution of selected regions from serum JRES spectrum. Panels show resonances generated by Valine (top), Leucine (middle) and Isoleucine (bottom). On the x-axis we report the chemical shift region of the multiplet. On the y-axis we report the intensity of the multiplet. The data is vectorised columnwise and plotted in 2D.

A10 Sensitivity analysis

We perform a sensitivity analysis using the urine spectrum analysed in Section 4.6. We investigate the robustness of posterior inference on the three most important parameters (concentration β, chemical shift δ, J-coupling translation ζ) to the choice of the prior hyperparameters for λ (scalar precision), τ (global shrinkage) and μ_{ijl} (local shrinkage). We do not consider the sensitivity with respect to the remaining hyperparameters, since those are either well informed from expert knowledge and experimental conditions (e.g. the peak width parameter σ) or guided by the specific application (e.g. $c_l = 0$).

Scalar precision parameter: In Section 4.3, we present results for the scalar precision parameter λ being a Gamma random variable with shape parameter $a = 10^{-6}$ and rate $b/2 = 10^{-9}/2$ (mean 2×10^3 and variance 4×10^{12}). We compare inference obtained using this prior with those obtained for two different hyperparameter choices: (i) Shape parameter $a = 10^{-7}$ and rate $b/2 = 10^{-10}/2$ (mean 2×10^3 and variance 4×10^{13}); (ii) Shape parameter $a = 10^{-4}$ and rate $b/2 = 10^{-8}/2$ (mean 2×10^4 and variance 4×10^{12}). In other words, we investigate changes in inference caused by a different choice of the mean of the prior distribution or of the prior variance. From Figures 14 and 15, we conclude that when only changing the variance of the prior distribution of λ, the posterior estimate of the shift parameters remains almost unchanged. On the other hand, changing the mean of the prior distribution of λ, we obtain different posterior results for the shift parameters. In our simulations, increasing the mean of the prior distribution of λ results in underestimation of the concentration parameters of some metabolites (Alanine and Lactate), see Figure 15. In short, it is better to choose a proper prior distribution for λ according to experimental noise. In our experience, a Gamma distribution with shape parameter $a = 10^{-6}$ and rate $b/2 = 10^{-9}/2$ works well in both 1D and 2D data analysis.

Global shrinkage parameter: In Section 4.3, we present results when the global shrinkage parameter τ follows a half Cauchy distribution $C^+(0,d)$ with $d = 10^{3.5}$. To investigate sensitivity, we perform posterior inference also for $d = 10^{2.5}$ and $10^{4.5}$. From Figures 16 and 17, we conclude that increasing d from $10^{3.5}$ to $10^{4.5}$ results in a slight change in the posterior estimates of both δ and ζ, while the estimation of the concentration parameters are robust. Decreasing d from $10^{3.5}$ to $10^{2.5}$ results in an overestimation of the concentration parameters, as well as an increase in the posterior estimate of the mean squared error, due to the fact that the framelet component of the model is able to accommodate for the overestimation of metabolite concentrations by concentrating the posterior reconstruction of the uncatalogued signal around negative values. Since the concentration is usually the main parameter of interest, it is advisable to start the analysis with a large value of d.

Local shrinkage parameters: The choice of the half Cauchy distribution for the local shrinkage parameters μ_{ijl} is discussed in detail in Section 4.3. The hyperparameter c_h in the prior distribution of the μ_{ijl} controls the strength of local shrinkage and in the main manuscript we fix $c_h = 5$. Here, we explore the sensitivity with respect to c_h, considering also the values $c_h = 4$ and 6. From Figures 18 and 19, it can be seen that increasing c_h affects the posterior distribution of ζ, while decreasing c_h leads to different inference for δ. Posterior estimates of the concentration parameters β_m also change slightly for different values of c_h.

Finally, we conclude that posterior inference is more sensitive to the choice of c_h than to the choice of λ or τ, and extra caution needs to be taken when setting c_h. The choice usually depends on the amount of overlap with uncatalogued signals and the presence or absence of isolated multiplets for each metabolite template.

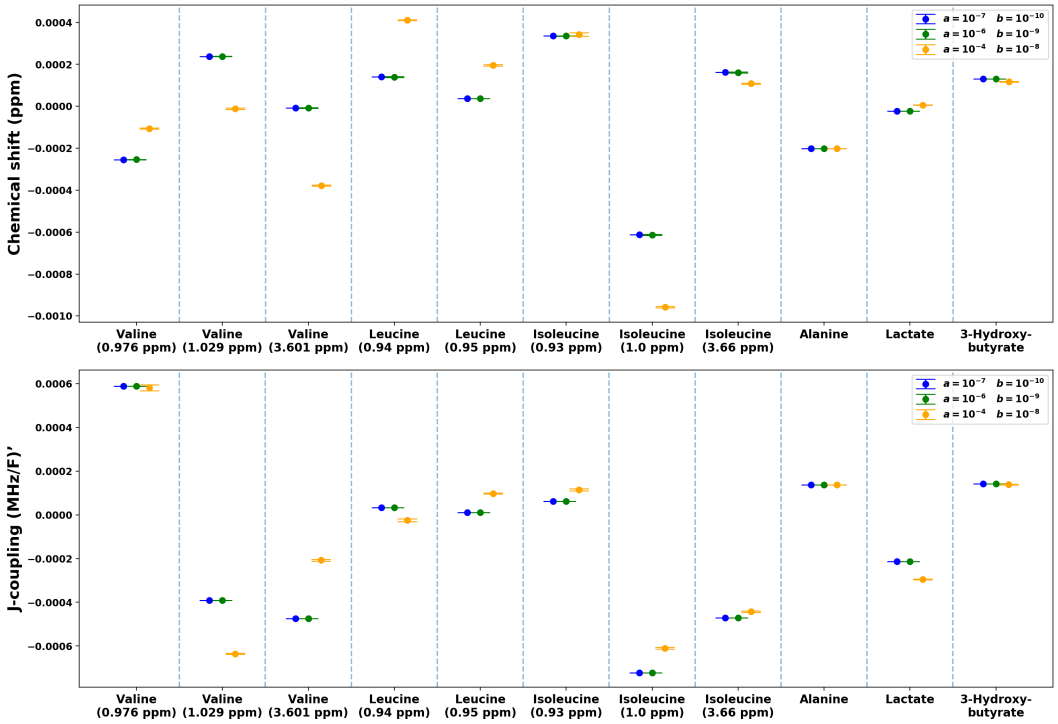

Figure 14: Comparison of posterior means of the shift in peak locations obtained with different prior distributions for the scalar precision parameter λ.

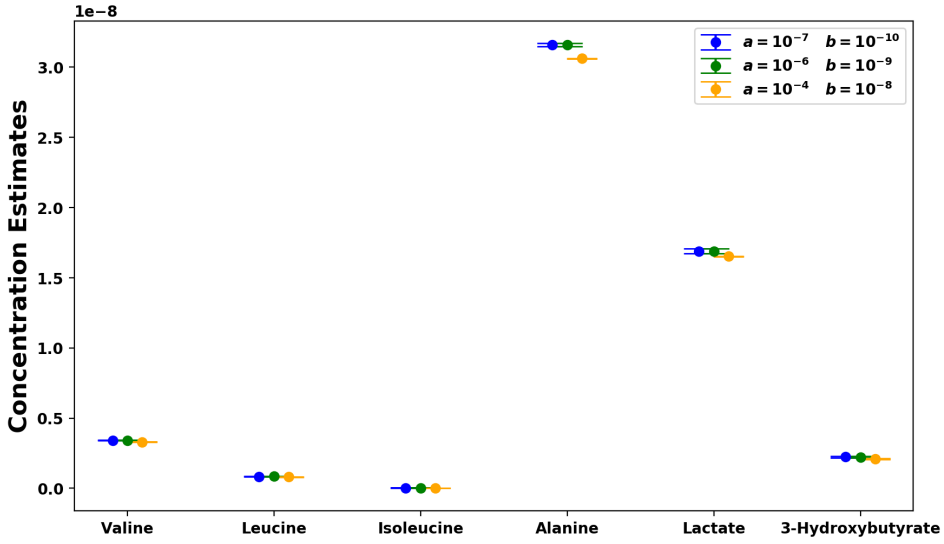

Figure 15: Comparison of posterior means of concentration parameters obtained with different prior distributions for the scalar precision parameter λ.

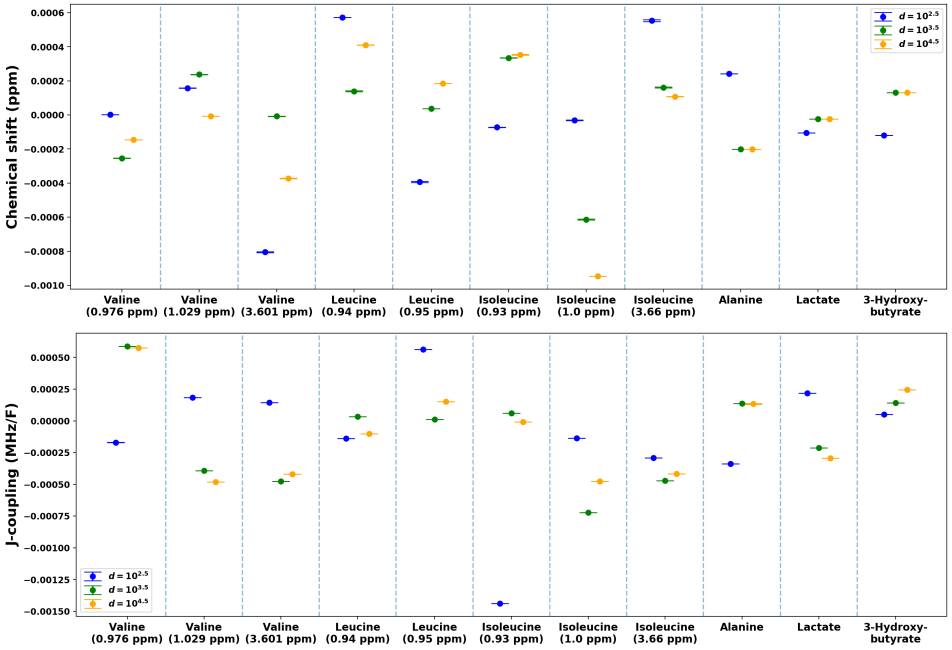

Figure 16: Comparison of shift in peak locations with different prior distributions for the global shrinkage parameter τ.

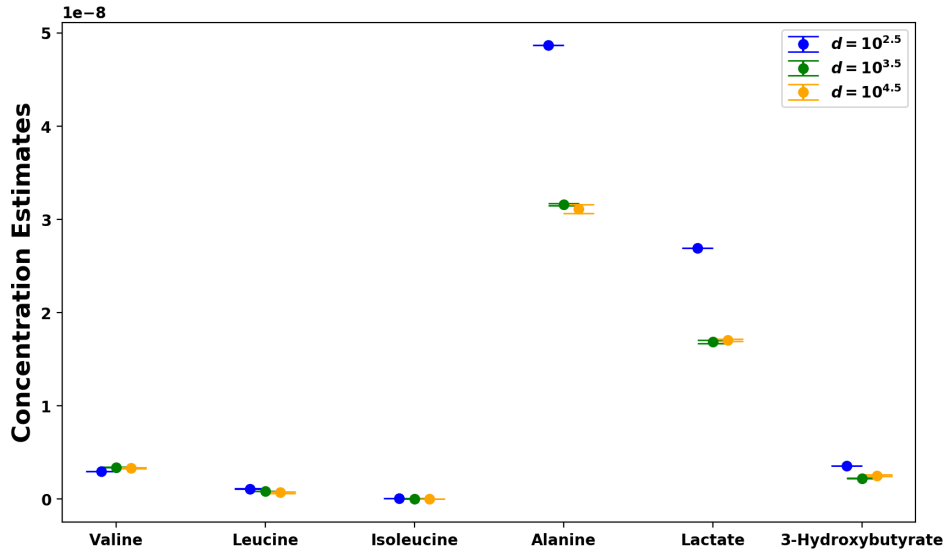

Figure 17: Comparison of estimated concentration with different prior distributions for the global shrinkage parameter τ.

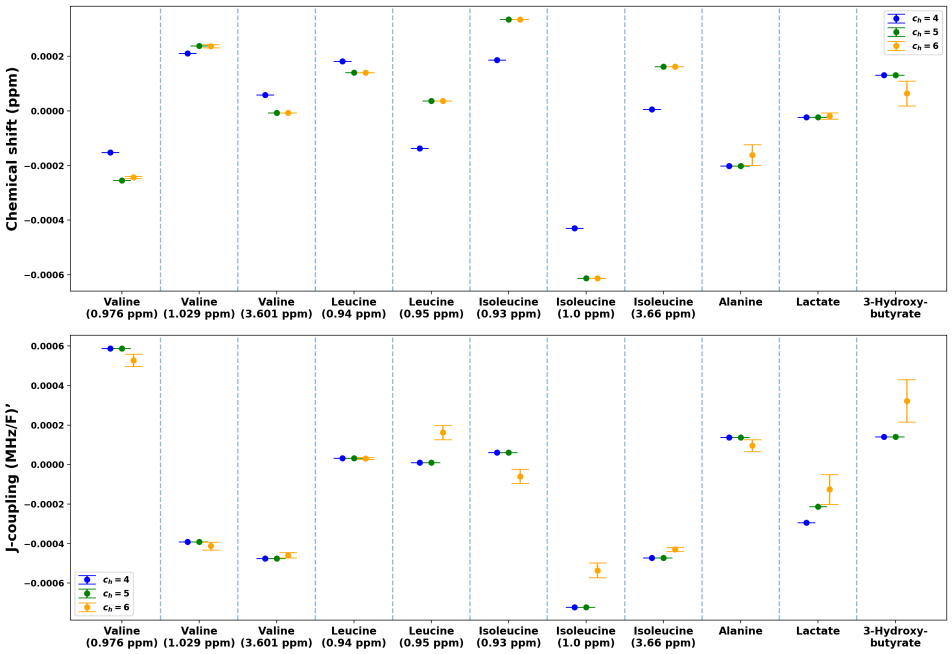

Figure 18: Comparison of posterior estimates of shift in peak locations obtained with different prior distributions for the local shrinkage parameters μ_{ijl}.

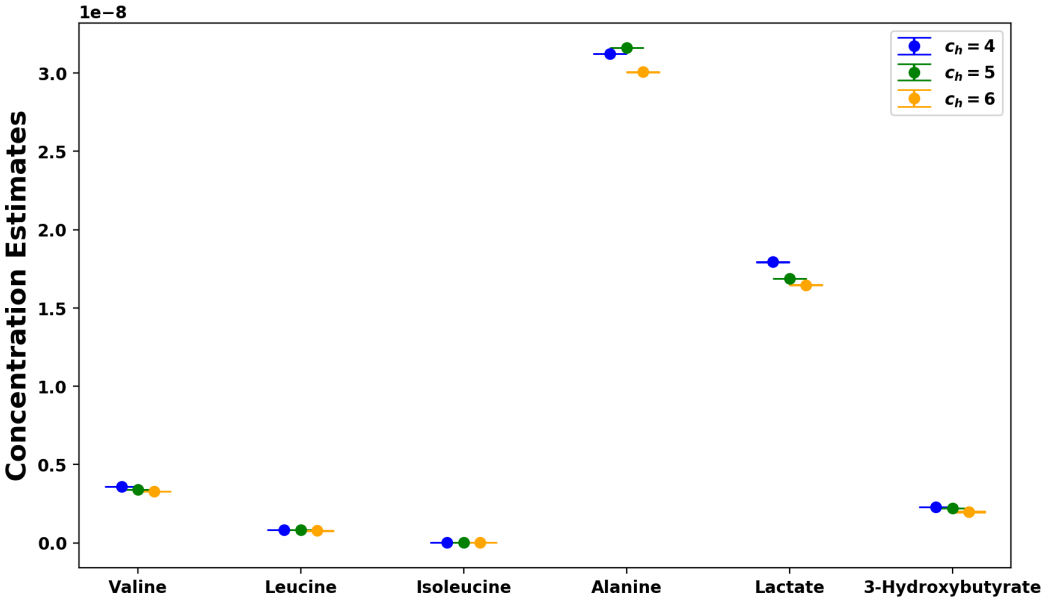

Figure 19: Comparison of posterior estimates of concentration with different prior distributions for the local shrinkage parameters μ_{ijl}.

A11 Posterior distributions for serum and urine spectra

Here we include some posterior distributions of concentration parameters, chemical shift parameters, and J-coupling parameters for the serum spectra discussed in Section A9 above (Figures 20, 21, 22) and for the urine spectra (Figures 23, 24, 25) discussed in Section 4.6.1. Due to peak overlap and shift, almost all posteriors are multi-modal distributions. This corresponds to the challenges met in convergence.

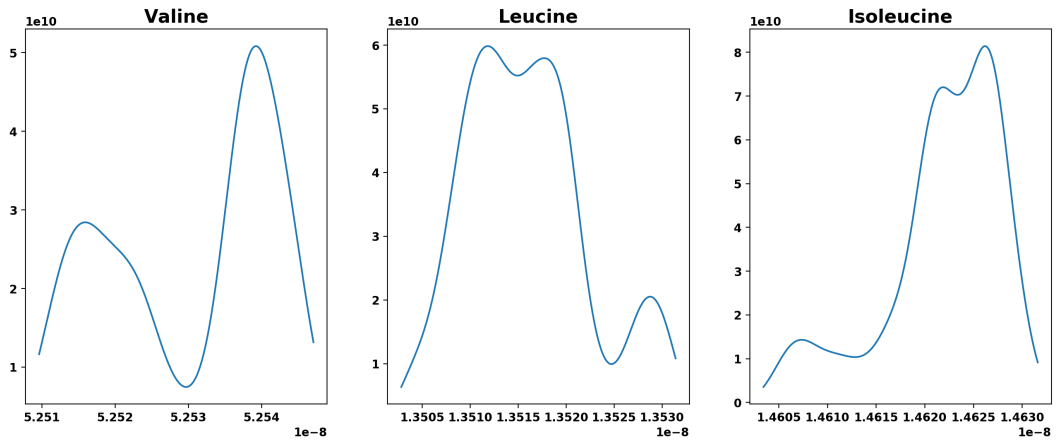

Figure 20: Posterior distributions of concentration parameters of the serum spectra.

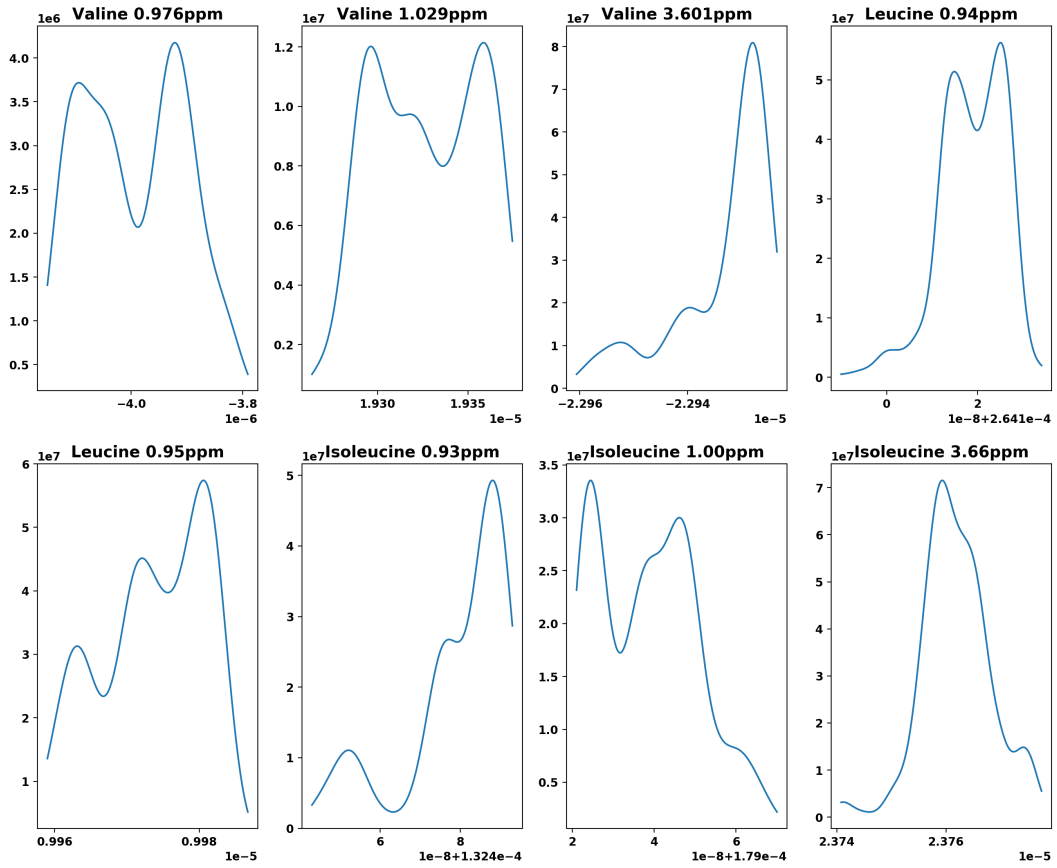

Figure 21: Posterior distributions of chemical shift parameters of the serum spectra.

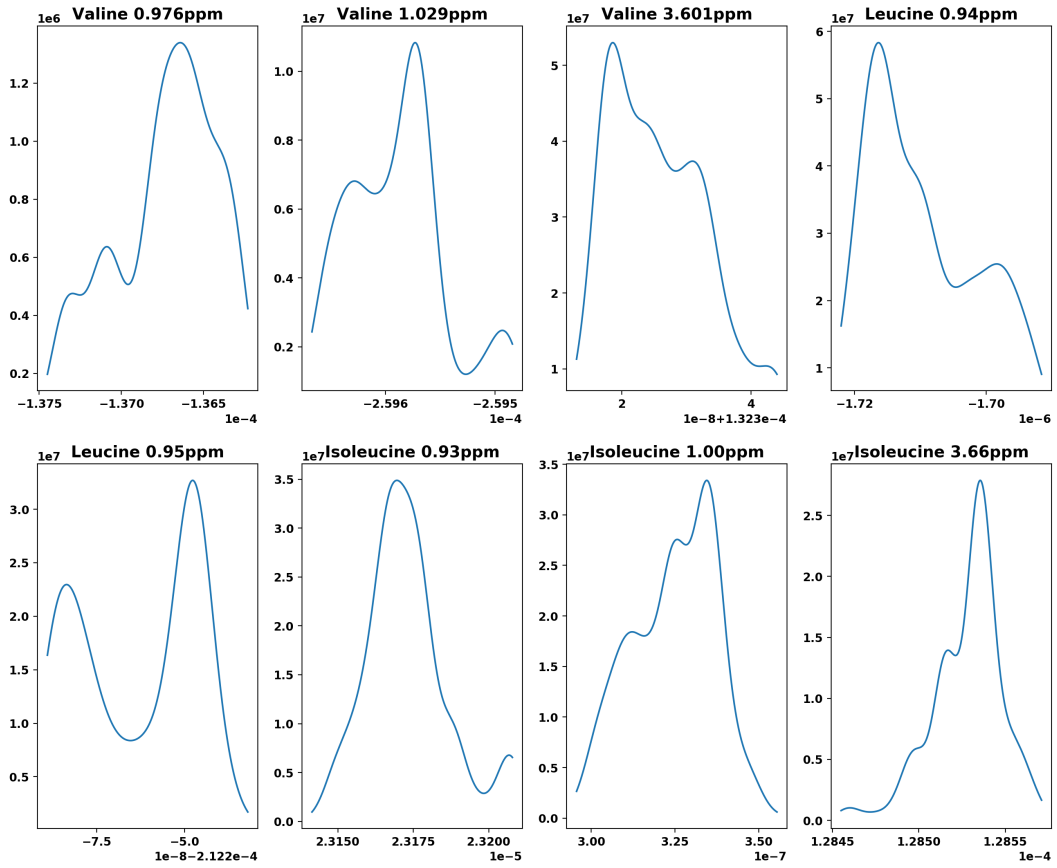

Figure 22: Posterior distribution of the J-coupling parameters of the serum spectra.

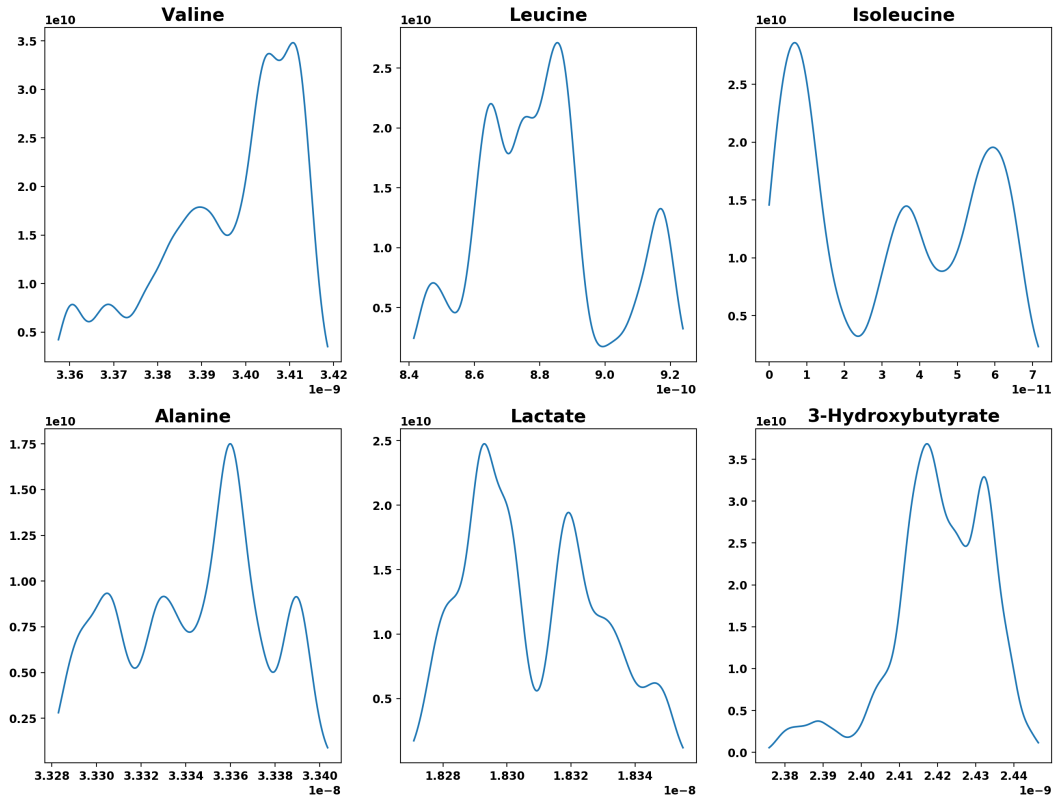

Figure 23: Posterior distributions of the concentration parameters of the urine spectra.

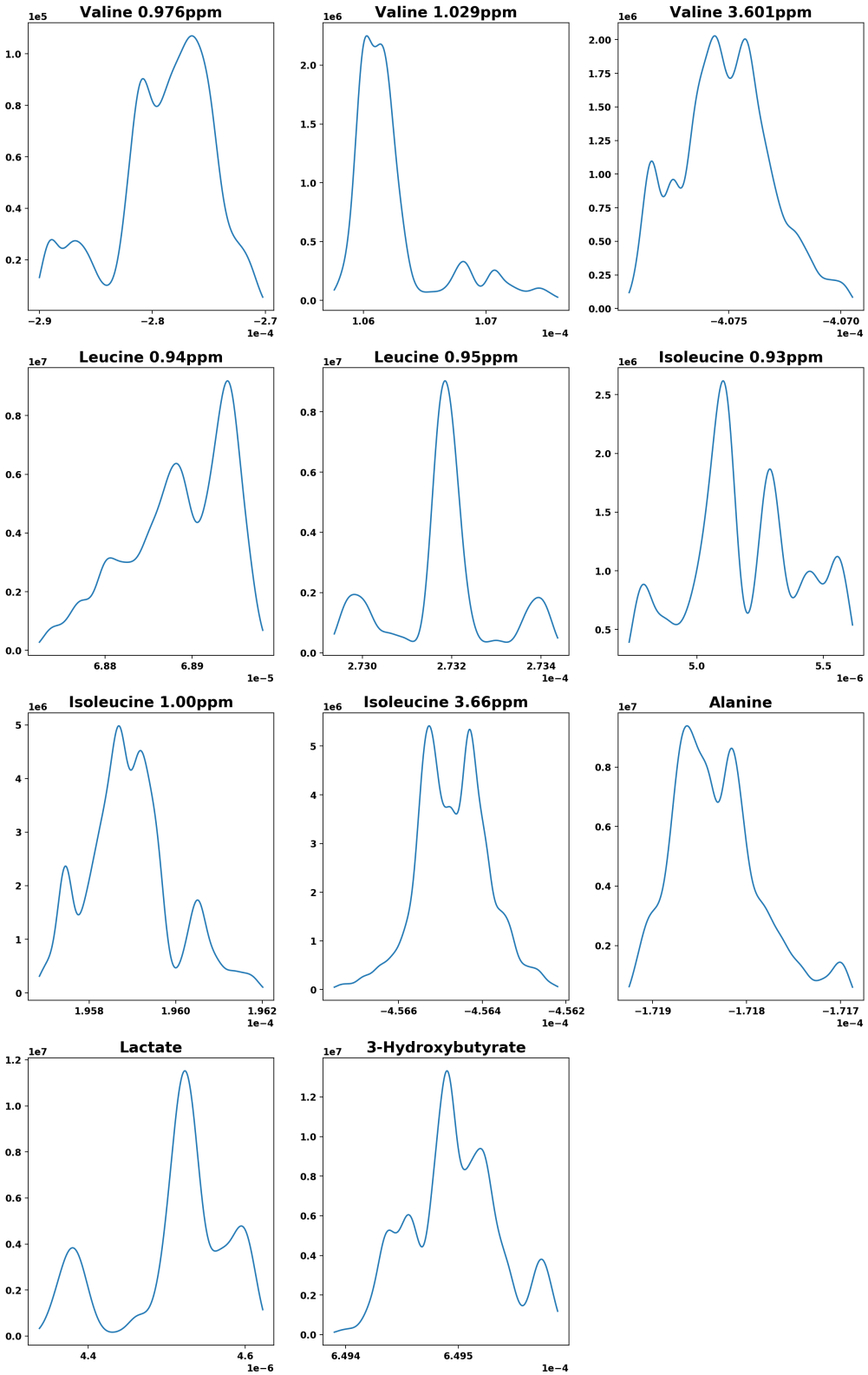

Figure 24: Posterior distributions of the chemical shift parameters of the urine spectra.

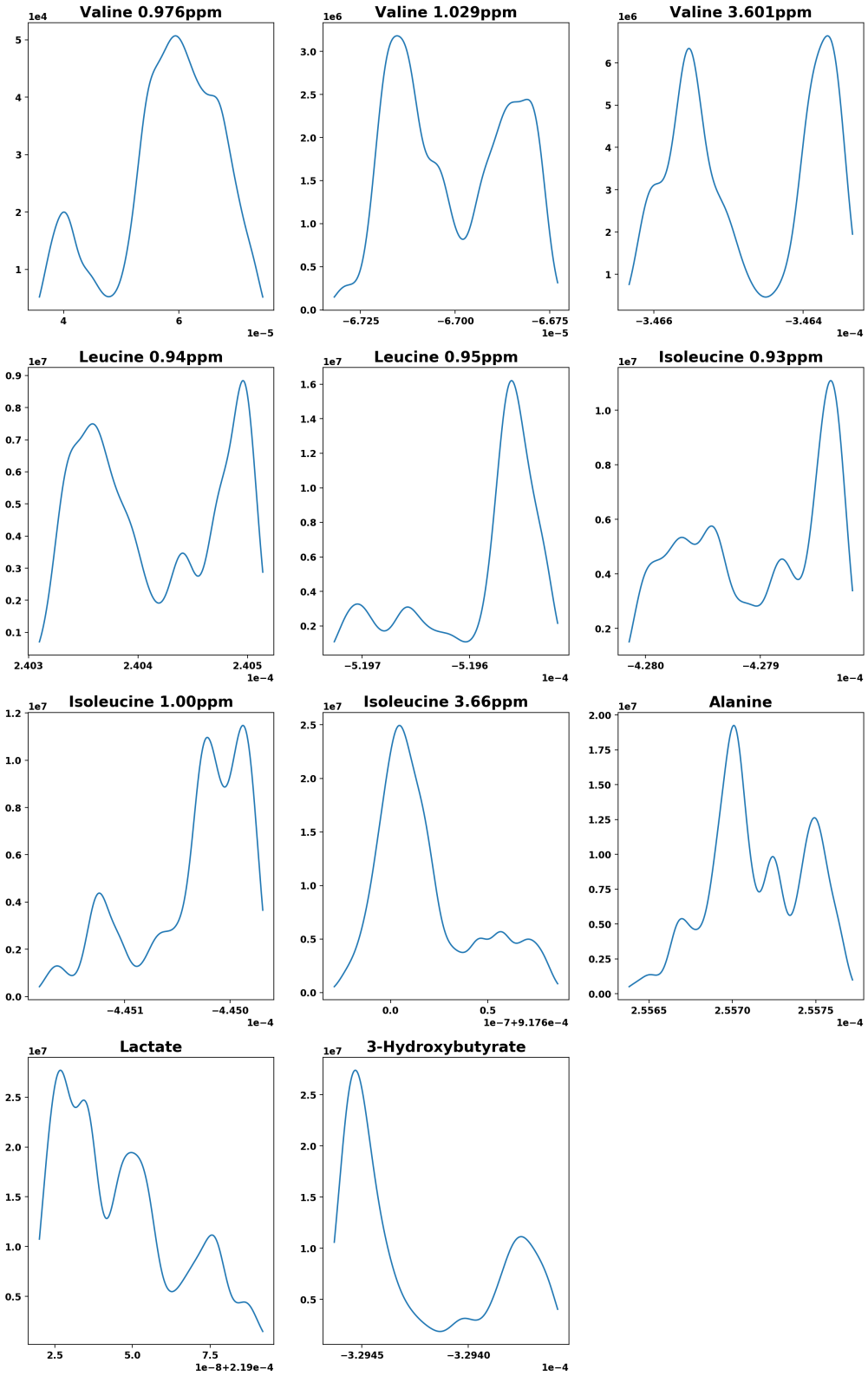

Figure 25: Posterior distribution of the *J*-coupling parameters of the urine spectra.

Bibliography

[1] Alquier, P. and Lounici, K. (2011). Pac-bayesian bounds for sparse regression estimation with exponential weights. *Electron. J. Statist.*, 5:127–145.

[2] Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

[3] Astle, W., De Iorio, M., Richardson, S., Stephens, D., and Ebbels, T. (2012). A Bayesian model of NMR spectra for the deconvolution and quantification of metabolites in complex biological mixtures. *Journal of the American Statistical Association*, 107(500):1259–1271.

[4] Aue, W. P., Bartholdi, E., and Ernst, R. R. (1976a). Two-dimensional spectroscopy. application to nuclear magnetic resonance. *The Journal of Chemical Physics*, 64(5):2229–2246.

[5] Aue, W. P., Karhan, J., and Ernst, R. R. (1976b). Homonuclear broad band decoupling and two-dimensional J-resolved NMR spectroscopy. *The Journal of Chemical Physics*, 64(10):4226–4227.

[6] Balazs, P., Dörfler, M., Jaillet, F., Holighaus, N., and Velasco, G. (2011). Theory, implementation and applications of nonstationary gabor frames. *Journal of Computational and Applied Mathematics*, 236(6):1481 – 1496.

[7] Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

[8] Bertsekas, D. P. (1999). *Nonlinear programming*. Athena scientific Belmont.

[9] Bhadra, A., Datta, J., Polson, N. G., Willard, B., et al. (2019). Lasso meets horseshoe: A survey. *Statistical Science*, 34(3):405–427.

[10] Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–laplace priors for optimal shrinkage. *J Amer Stat Assoc*, 110(512):1479–1490.

[11] Bieleń, A., Mrochem-Kwarciak, J., Skorupa, A., Ciszek, M., Heyda, A., Wygoda, A., Kotylak, A., Składowski, K., Sokół, M., et al. (2019). Nmr-based metabolomics in real-time monitoring of treatment induced toxicity and cachexia in head and neck cancer: a method for early detection of high risk patients. *Metabolomics*, 15(8):110.

[12] Bingol, K., Zhang, F., Bruschweiler-Li, L., and Brüschweiler, R. (2013). Quantitative analysis of metabolic mixtures by two-dimensional 13c constant-time tocsy nmr spectroscopy. *Analytical chemistry*, 85(13):6414–6420.

[13] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

[14] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

[15] Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):265–285.

[16] Bottou, L. (2012). *Stochastic Gradient Descent Tricks*, pages 421–436. Springer Berlin Heidelberg, Berlin, Heidelberg.

[17] Braunschweiler, L. and Ernst, R. (1983). Coherence transfer by isotropic mixing: Application to proton correlation spectroscopy. *Journal of Magnetic Resonance*, 53(3):521 – 528.

[18] Brindle, J. T., Antti, H., Holmes, E., Tranter, G., Nicholson, J. K., Bethell, H. W. L., Clarke, S., Schofield, P. M., McKilligin, E., Mosedale, D. E., and Grainger, D. J. (2002). Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. *Nature Medicine*, 8:1439 EP –.

[19] Bruce, S. D., Higinbotham, J., Marshall, I., and Beswick, P. H. (2000). An analytical derivation of a popular approximation of the voigt function for quantification of nmr spectra. *Journal of Magnetic Resonance*, 142(1):57–63.

[20] Bundy, J. G., Spurgeon, D. J., Svendsen, C., Hankard, P. K., Osborn, D., Lindon, J. C., and Nicholson, J. K. (2002). Earthworm species of the genus Eisenia can be phenotypically differentiated by metabolic profiling. *FEBS Letters*, 521(1):115 – 120.

[21] Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1–32.

[22] Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

[23] Casazza, P. G. and Kutyniok, G. (2012). *Finite Frames: Theory and Applications*. Birkhäuser Basel.

[24] Casella, G. and Robert, C. P. (1996). Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94.

[25] Cavanagh, J., Skelton, N., Fairbrother, W., Rance, M., Palmer, A., Skelton, N., and Rance, M. (2007). *Protein NMR Spectroscopy: Principles and Practice*. Academic Press.

[26] Chan, K. S. and Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving counts. *Journal of the American Statistical Association*, 90(429):242–252.

[27] Chen, C., Gonzalez, F. J., and Idle, J. R. (2007). Lc-ms-based metabolomics in drug metabolism. *Drug Metabolism Reviews*, 39(2-3):581–597. PMID: 17786640.

[28] Christen, J. A. and Fox, C. (2010). A general purpose sampling algorithm for continuous distributions (the t -walk). *Bayesian Anal.*, 5(2):263–281.

[29] Cowles, M. K. and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

[30] Craig, A., Cloarec, O., Holmes, E., Nicholson, J., and Lindon, J. (2006). Scaling and normalization effects in nmr spectroscopic metabonomic data sets. *Analytical Chemistry*, 78(7):2262–7.

[31] Daubechies, I., Grossmann, A., and Meyer, Y. (1986). Painless nonorthogonal expansions. *Journal of Mathematical Physics*, 27.

[32] Davis, D. G. and Bax, A. (1985). Assignment of complex proton nmr spectra via two-dimensional homonuclear hartmann-hahn spectroscopy. *Journal of the American Chemical Society*, 107(9):2820–2821.

[33] Defernez, M. and Colquhoun, I. J. (2003). Factors affecting the robustness of metabolite fingerprinting using 1h nmr spectra. *Phytochemistry*, 62(6):1009 – 1017. Plant Metabolomics.

[34] Dehghan, A. (2019). Linking metabolic phenotyping and genomic information. In *The Handbook of Metabolic Phenotyping*, pages 561–569. Elsevier.

[35] Dettmer, K., Aronov, P. A., and Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1):51–78.

[36] Dona, A. C., Jiménez, B., Schäfer, H., Humpfer, E., Spraul, M., Lewis, M. R., Pearce, J. T. M., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2014). Precision high-throughput proton nmr spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Analytical Chemistry*, 86(19):9887–9894.

[37] Dong, B. and Shen, Z. (2010). MRA-based wavelet frames and applications. *IAS/Park City Mathematics Series*, 19.

[38] Dong, B. and Shen, Z. (2015). Image restoration: a data-driven perspective. In *Proceedings of the International Congress on Industrial and Applied Mathematics (ICIAM)*, pages 65–108, Beijing, China. High Education Press.

[39] Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

[40] Duffin, R. and Schaeffer, A. (1952). A class of nonharmonic Fourier series. *Transactions of the American Mathematical Society*, 72:341–366.

[41] Dufour, G., Evrard, B., and de Tullio, P. (2015). 2d-cosy nmr spectroscopy as a quantitative tool in biological matrix: Application to cyclodextrins. *The AAPS Journal*, 17(6):1501–1510.

[42] Ebbels, T. M. and Cavill, R. (2009). Bioinformatic methods in nmr-based metabolic profiling. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 55(4):361–374.

[43] Elliott, P., Vergnaud, A.-C., Singh, D., Neasham, D., Spear, J., and Heard, A. (2014). The airwave health monitoring study of police officers and staff in great britain: Rationale, design and methods. *Environmental Research*, 134:280 – 285. Linking Exposure and Health in Environmental Public Health Tracking.

[44] Everitt, B. S. (2006). *The Cambridge dictionary of statistics; 3rd ed.* Cambridge University Press, Cambridge.

[45] Fan, T. W.-M. (1996). Metabolite profiling by one- and two-dimensional nmr analysis of complex mixtures. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 28(2):161 – 219.

[46] Féraud, B., Govaerts, B., Verleysen, M., and Tullio, P. (2015). Statistical treatment of 2D NMR COSY spectra in metabolomics: data preparation, clustering-based evaluation of the metabolomic informative content and comparison with H-NMR. *Metabolomics*, 11(6):1756 – 1768.

[47] Fonville, J. M., Maher, A. D., Coen, M., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2010). Evaluation of full-resolution j-resolved 1h nmr projections of biofluids for metabonomics information retrieval and biomarker identification. *Analytical Chemistry*, 82(5):1811–1821. PMID: 20131799.

[48] Forbes, F. and Fort, G. (2007). Combining Monte Carlo and mean-field-like methods for inference in hidden Markov random fields. *IEEE Transactions on Image Processing*, 16(3):824–837.

[49] Forgacs, A. L., Kent, M. N., Makley, M. K., Mets, B., DelRaso, N., Jahns, G. L., Burgoon, L. D., Zacharewski, T. R., and Reo, N. V. (2011). Comparative metabolomic and genomic analyses of TCDD-elicited metabolic disruption in mouse and rat liver. *Toxicological Sciences*, 125(1):41–55.

[50] Fort, G., Moulines, E., et al. (2003). Convergence of the monte carlo expectation maximization for curved exponential families. *The Annals of Statistics*, 31(4):1220–1259.

[51] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.*, 1(3):515–534.

[52] George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

[53] Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. (2009). Pac-bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 353–360, New York, NY, USA. ACM.

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

[55] Gómez, J., Brezmes, J., Mallol, R., Rodríguez, M. A., Vinaixa, M., Salek, R. M., Correig, X., and Cañellas, N. (2014). Dolphin: a tool for automatic targeted metabolite profiling using 1d and 2d 1h-nmr data. *Analytical and Bioanalytical Chemistry*, 406(30):7967–7976.

[56] Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., and Kell, D. B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in Biotechnology*, 22(5):245 – 252.

[57] Griffiths, J. R., McSheehy, P. M. J., Robinson, S. P., Troy, H., Chung, Y. L., Leek, R., Williams, K. J., Stratford, I. J., Harris, A. L., and Stubbs, M. (2002). Metabolic changes detected by in vivo magnetic resonance studies of HEPA-1

wild-type tumors and tumors deficient in hypoxia-inducible factor-1beta (HIF-1beta): evidence of an anabolic role for the HIF-1 pathway. *Cancer research*, 62 3:688–95.

[58] H., A. J. J., E., S. G., M., S. W., Zehua, Z., and L., E. J. The nmr chemical shift ph measurement revisited: Analysis of error and modeling of a ph dependent reference. *Magnetic Resonance in Medicine*, 36(5):674–683.

[59] Hajduk, A., Mrochem-Kwarciak, J., Skorupa, A., Ciszek, M., Heyda, A., Składowski, K., Sokół, M., et al. (2016). 1 h nmr based metabolomic approach to monitoring of the head and neck cancer treatment toxicity. *Metabolomics*, 12(6):102.

[60] Hao, J., Astle, W., De Iorio, M., and Ebbels, T. M. D. (2012). Batman: an r package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a bayesian model. *Bioinformatics*, 28(15):2088–2090.

[61] Hao, J., Liebeke, M., Astle, W., De Iorio, M., Bundy, J., and Ebbels, T. M. D. (2014). Bayesina deconvolution and quantification of metabolites in compex 1D NMR spectra using batman. *Nature Protocols*, 9(6):1416.

[62] Helmus, J. J. and Jaroniec, C. P. (2013). Nmrglue: an open source python package for the analysis of multidimensional nmr data. *Journal of biomolecular NMR*, 55(4):355–367.

[63] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.

[64] Hollinshead, K. E., Williams, D. S., Tennant, D. A., and Ludwig, C. (2016). Probing cancer cell metabolism using nmr spectroscopy. In *Tumor Microenvironment*, pages 89–111. Springer.

[65] Holmes, E., Loo, R. L., Stamler, J., Bictash, M., Yap, I. K. S., Chan, Q., Ebbels, T., De Iorio, M., Brown, I. J., Veselkov, K. A., Daviglus, M. L., Kesteloot, H., Ueshima, H., Zhao, L., Nicholson, J. K., and Elliott, P. (2008a). Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature*, 453:396 EP.

[66] Holmes, E., Loo, R. L., Stamler, J., Bictash, M., Yap, I. K. S., Chan, Q., Ebbels, T., De Iorio, M., Brown, I. J., Veselkov, K. A., Daviglus, M. L., Kesteloot, H., Ueshima, H., Zhao, L., Nicholson, J. K., and Elliott, P. (2008b). Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature*, 453:396 EP –.

[67] Hore, P. (2015a). *Nuclear Magnetic Resonance*. Oxford chemistry primers. Oxford University Press.

[68] Hore, P. J. (2015b). *Nuclear Magnetic Resonance*. Oxford University Press.

[69] Ibrahim, S. M. and Gold, R. (2005). Genomics, proteomics, metabolomics: what is in a word for multiple sclerosis? *Current Opinion in Neurology*, 18(3):231–235.

[70] Illig, T., Gieger, C., Zhai, G., Römisch-Margl, W., Wang-Sattler, R., Prehn, C., Altmaier, E., Kastenmüller, G., Kato, B. S., Mewes, H.-W., Meitinger, T., de Angelis, M. H., Kronenberg, F., Soranzo, N., Wichmann, H.-E., Spector, T. D., Adamski, J., and Suhre, K. (2009). A genome-wide perspective of genetic variation in human metabolism. *Nature Genetics*, 42:137 EP –.

[71] Jeener, J. (2007). *Jeener, Jean: Reminiscences about the Early Days of 2D NMR*. American Cancer Society.

[72] Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An

introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

[73] Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing*, 21:93–105.

[74] Karakach, T. K., Wentzell, P. D., and Walter, J. A. (2009). Characterization of the measurement error structure in 1d 1h nmr data for metabolomics studies. *Analytica Chimica Acta*, 636(2):163 – 174.

[75] Khalil, I. G. and Hill, C. (2005). Systems biology for cancer. *Current Opinion in Oncology*, 17(1).

[76] Kikuchi, J., Tsuboi, Y., Komatsu, K., Gomi, M., Chikayama, E., and Date, Y. (2016). Spincouple: Development of a web tool for analyzing metabolite mixtures via two-dimensional j-resolved nmr database. *Analytical Chemistry*, 88(1):659–665. PMID: 26624790.

[77] Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474.

[78] Levine, R. A. and Casella, G. (2001). Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439.

[79] Li, Y., Hernández-Lobato, J. M., and Turner, R. E. (2015). Stochastic expectation propagation. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 2323–2331. Curran Associates, Inc.

[80] Lindon, J. C. and Nicholson, J. K. (2008). Analytical technologies for metabonomics and metabolomics, and multi-omic information recovery. *TrAC Trends in Analytical Chemistry*, 27(3):194 – 204. Metabolomics.

[81] Lindon, J. C., Nicholson, J. K., Holmes, E., Antti, H., Bollard, M. E., Keun, H., Beckonert, O., Ebbels, T. M., Reily, M. D., Robertson, D., Stevens, G. J., Luke, P., Breau, A. P., Cantor, G. H., Bible, R. H., Niederhauser, U., Senn, H., Schlotterbeck, G., Sidelmann, U. G., Laursen, S. M., Tymiak, A., Car, B. D., Lehman-McKeeman, L., Colet, J.-M., Loukaci, A., and Thomas, C. (2003). Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. *Toxicology and Applied Pharmacology*, 187(3):137 – 146.

[82] Lindon, J. C., Nicholson, J. K., Holmes, E., and Everett, J. R. (2000). Metabonomics: Metabolic processes studied by nmr spectroscopy of biofluids. *Concepts in Magnetic Resonance*, 12(5):289–320.

[83] Link, W. A. and Eaton, M. J. (2012). On thinning of chains in mcmc. *Methods in Ecology and Evolution*, 3(1):112–115.

[84] Ludwig, C. and Viant, M. R. (2010). Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochemical Analysis*, 21(1):22–32.

[85] Lundblad, R. and Macdonald, F. (2010). *Handbook of Biochemistry and Molecular Biology, Fourth Edition*. Taylor & Francis.

[86] Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs - a bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337.

[87] Maceachern, S. N. and Berliner, L. M. (1994). Subsampling the gibbs sampler. *The American Statistician*, 48(3):188–190.

[88] Mahrous, E. A. and Farag, M. A. (2015). Two dimensional nmr spectroscopic

approaches for exploring plant metabolome: a review. *Journal of advanced research*, 6(1):3–15.

[89] Mallat, S. (2008). *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, Inc.

[90] Martino, L. and Míguez, J. (2011). A generalization of the adaptive rejection sampling algorithm. *Statistics and Computing*, 21(4):633–647.

[91] Minka, T. P. (2001). Expectation propagation for approximate bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pages 362–369, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[92] Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

[93] Moore, G. J. and Sillerud, L. O. (1994). The pH Dependence of Chemical Shift and Spin-Spin Coupling for Citrate. *Journal of Magnetic Resonance*, 103:87–88.

[94] Nicholson, J. K. (2006). Global systems biology, personalized medicine and molecular epidemiology. *Molecular Systems Biology*, 2(1).

[95] Nicholson, J. K., Connelly, J., Lindon, J. C., and Holmes, E. (2002). Metabonomics: a platform for studying drug toxicity and gene function. *Nature Reviews Drug Discovery*, 1:153 EP.

[96] Nicholson, J. K., Lindon, J. C., and Holmes, E. (1999). 'metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological nmr spectroscopic data. *Xenobiotica*, 29(11):1181–1189. PMID: 10598751.

[97] Nikolsky, Y., Nikolskaya, T., and Bugrim, A. (2005). Biological networks and analysis of experimental data in drug discovery. *Drug Discovery Today*, 10(9):653 – 662.

[98] Opper, M., Çakmak, B., and Winther, O. (2016). A theory of solving TAP equations for ising models with general invariant random matrices. *Journal of Physics A: Mathematical and Theoretical*, 49(11):114002.

[99] Palaric, C., Pilard, S., Fontaine, J.-X., Boccard, J., Mathiron, D., Rigaud, S., Cailleu, D., Mesnard, F., Gut, Y., Renaud, T., et al. (2019). Processing of nmr and ms metabolomics data using chemometrics methods: a global tool for fungi biotransformation reactions monitoring. *Metabolomics*, 15(8):107.

[100] Parsons, H. M., Ludwig, C., Günther, U. L., and Viant, M. R. (2007). Improved classification accuracy in 1-and 2-dimensional nmr metabolomics data using the variance stabilising generalised logarithm transformation. *BMC bioinformatics*, 8(1):234.

[101] Piironen, J., Vehtari, A., et al. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic J. Stat.*, 11(2):5018–5051.

[102] Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling.

[103] Polson, N. G., Scott, J. G., et al. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902.

[104] Raamsdonk, L. M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M. C., Berden, J. A., Brindle, K. M., Kell, D. B., Rowland, J. J., Westerhoff, H. V., van Dam, K., and Oliver, S. G. (2001). A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology*, 19:45 EP –.

[105] Ranganath, R., Gerrish, S., and Blei, D. (2014). Black Box Variational Inference. In Kaski, S. and Corander, J., editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland. PMLR.

[106] Ripley, B. (2009). *Stochastic Simulation*. Wiley Series in Probability and Statistics. Wiley.

[107] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407.

[108] Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

[109] Robert, C. P. and Casella, G. (1999). *The Metropolis—Hastings Algorithm*, pages 231–283. Springer New York, New York, NY.

[110] Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367.

[111] Robertson, D. G., Reily, M. D., Sigler, R. E., Wells, D. F., Paterson, D. A., and Braden, T. K. (2000). Metabonomics: Evaluation of nuclear magnetic resonance (nmr) and pattern recognition technology for rapid in vivo screening of liver and kidney toxicants. *Toxicological Sciences*, 57(2):326–337.

[112] Ron, A. and Shen, Z. (1997). Affine systems in $L_2(\mathbb{R}^d)$: the analysis of the analysis operator. *Journal of Functional Analysis*, 148:408–447.

[113] Ross, S. M. (2002). *Simulation*. Elsevier.

[114] Sandusky, P. and Raftery, D. (2005). Use of selective tocsy nmr experiments for quantifying minor components in complex mixtures: Application to

the metabonomics of amino acids in honey. *Analytical Chemistry*, 77(8):2455–2463.

[115] Shulaev, V. (2006). Metabolomics technology and bioinformatics. *Briefings in Bioinformatics*, 7(2):128–139.

[116] Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765.

[117] Sobolev, A. P., Brosio, E., Gianferri, R., and Segre, A. L. (2005). Metabolic profile of lettuce leaves by high-field nmr spectra. *Magnetic Resonance in Chemistry*, 43(8):625–638.

[118] Sousa, S., Magalhães, A., and Ferreira, M. M. C. (2013). Optimized bucketing for nmr spectra: Three case studies. *Chemometrics and Intelligent Laboratory Systems*, 122:93–102.

[119] Swaroop, S. and Turner, R. E. (2017). Understanding expectation propagation. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[120] Szakacs, Z., Hagele, G., and Tyka, R. (2004). 1h/31p nmr ph indicator series to eliminate the glass electrode in nmr spectroscopic pka determinations. *Analytica Chimica Acta*, 522(2):247 – 258.

[121] Takis, P. G., Schäfer, H., Spraul, M., and Luchinat, C. (2017). Deconvoluting interrelationships between concentrations and chemical shifts in urine provides a powerful analysis tool. *Nature Communications*, 8(1):1662.

[122] ter Kuile, B. H. and Westerhoff, H. V. (2001). Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Letters*, 500(3):169–171.

[123] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

[124] Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1:211–244.

[125] Tran, M.-N., Nguyen, D. H., and Nguyen, D. (2019). Variational Bayes on manifolds. arxiv preprint arXiv:1908.03097.

[126] Tran, M.-N., Nott, D. J., Kuk, A. Y., and Kohn, R. (2016). Parallel variational Bayes for large datasets with an application to generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 25(2):626–646.

[127] Tredwell, G. D., Bundy, J. G., De Iorio, M., and Ebbels, T. M. D. (2016). Modelling the acid/base 1h nmr chemical shift limits of metabolites in human urine. *Metabolomics*, 12(10):152.

[128] Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C. F., Tolmie, D. E., Kent Wenger, R., Yao, H., and Markley, J. L. (2007). BioMagResBank. *Nucleic Acids Research*, 36(suppl-1):D402–D408.

[129] Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C. F., Tolmie, D. E., Kent Wenger, R., Yao, H., and Markley, J. L. (2008). Biomagresbank. *Nucleic Acids Research*, 36:D402–D408.

[130] Viant, M. R. (2003). Improved methods for the acquisition and interpretation of NMR metabolomic data. *Biochemical and Biophysical Research Communications*, 310(3):943 – 948.

[131] Viswan, A., Singh, C., Kayastha, A. M., Azim, A., and Sinha, N. (2019). An nmr based panorama of the heterogeneous biology of acute respiratory distress syndrome (ards) from the standpoint of metabolic biomarkers. *NMR in Biomedicine*.

[132] Vu, T. N. and Laukens, K. (2013). Getting your peaks in line: a review of alignment methods for nmr spectral data. *Metabolites*, 3(2):259–276.

[133] Wainwright, M. J. and Jordan, M. I. (2014). *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers.

[134] Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704.

[135] Weljie, A. M., Newton, J., Mercier, P., Carlson, E., and Slupsky, C. M. (2006). Targeted profiling: Quantitative analysis of 1H NMR metabolomics data. *Analytical Chemistry*, 78(13):4430–4442.

[136] Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., Sayeeda, Z., Lo, E., Assempour, N., Berjanskii, M., Singhal, S., Arndt, D., Liang, Y., Badran, H., Grant, J., Serra-Cayuela, A., Liu, Y., Mandal, R., Neveu, V., Pon, A., Knox, C., Wilson, M., Manach, C., and Scalbert, A. (2018). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research*, 46(D1):D608–D617.

[137] Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R., and Scalbert, A. (2013). HMDB 3.0: The human metabolome database in 2013. *Nucleic Acids Research*, 41(D1):D801–D807.

[138] Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., Hau, D. D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J. A., Lim, E., Sobsey, C. A., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H. J., and Forsythe, I. (2009). HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Research*, 37:D603–D610.

[139] Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.-A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D. D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G. E., MacInnis, G. D., Weljie, A. M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B. D., Vogel, H. J., and Querengesser, L. (2007). HMDB: the human metabolome database. *Nucleic Acids Research*, 35:D521–D526.

[140] Ye, L., De Iorio, M., and Ebbels, T. M. D. (2018). Bayesian estimation of the number of protonation sites for urinary metabolites from nmr spectroscopic data. *Metabolomics*, 14(5):56.

[141] Zhang, S., Nagana Gowda, G. A., Ye, T., and Raftery, D. (2010). Advances in nmr-based biofluid analysis and metabolite profiling. *Analyst*, 135:1490–1498.