

# Bayesian Deconvolution and Quantification of Metabolites from $J$ -Resolved NMR Spectroscopy

Andreas Heinecke\*, Lifeng Ye†, Maria De Iorio‡, and Timothy Ebbels§

**Abstract.** Two-dimensional (2D) nuclear magnetic resonance (NMR) methods have become increasingly popular in metabolomics, since they have considerable potential to accurately identify and quantify metabolites within complex biological samples. 2D  $^1\text{H}$   $J$ -resolved (JRES) NMR spectroscopy is a widely used method that expands overlapping resonances into a second dimension. However, existing analytical processing methods do not fully exploit the information in the JRES spectrum and, more importantly, do not provide measures of uncertainty associated with the estimates of quantities of interest, such as metabolite concentration. Combining the data-generating mechanisms and the extensive prior knowledge available in online databases, we develop a Bayesian method to analyse 2D JRES data, which allows for automatic deconvolution, identification and quantification of metabolites. The model extends and improves previous work on one-dimensional NMR spectral data. Our approach is based on a combination of B-spline tight wavelet frames and theoretical templates, and thus enables the automatic incorporation of expert knowledge within the inferential framework. Posterior inference is performed through specially devised Markov chain Monte Carlo methods. We demonstrate the performance of our approach via analyses of datasets from serum and urine, showing the advantages of our proposed approach in terms of identification and quantification of metabolites.

**Keywords:** MCMC, metabolomics, NMR spectroscopy, shrinkage priors, wavelet frames.

## 1 Introduction

The metabolome is the collection of small biological molecules (metabolites) present in a living system. Metabolites are the building blocks of the large molecules of life (such as DNA or proteins) and also convey energy and information around the cell. Examples of metabolites include amino acids (e.g. Leucine), which form the building blocks of proteins, and glucose, which provides the energy to sustain life through the reactions of glycolysis. The scientific discipline concerned with the comprehensive quantitative analysis of metabolites is referred to as *metabolomics* (sometimes as metabonomics,

---

\*Yale-NUS College, Singapore 138527, Singapore, [andreas.heinecke@yale-nus.edu.sg](mailto:andreas.heinecke@yale-nus.edu.sg)

†Department of Statistical Sciences, University College, London WC1E 6BT, United Kingdom, [lifeng.ye.13@ucl.ac.uk](mailto:lifeng.ye.13@ucl.ac.uk)

‡Yale-NUS College, Singapore 138527, Singapore; Department of Statistical Sciences, University College, London WC1E 6BT, United Kingdom, [maria@yale-nus.edu.sg](mailto:maria@yale-nus.edu.sg)

§Computational and Systems Medicine, Department of Surgery and Cancer, Imperial College, London SW1 2AZ, United Kingdom, [t.ebbels@imperial.ac.uk](mailto:t.ebbels@imperial.ac.uk)

metabolic profiling or metabolic phenotyping). Almost all experiments in metabolomics require identification or quantification of metabolites in complex biological mixtures, usually biofluids or tissue samples. Research in nuclear magnetic resonance (NMR)-based metabolomics has obtained substantial attention in the biomedical sciences, with numerous applications in the areas of biology and medicine, including biochemistry (Raamsdonk et al., 2001; Palaric et al., 2019), oncology (Griffiths et al., 2002; Hollinshead et al., 2016), disease diagnostics (Brindle et al., 2002; Bieleń et al., 2019), epidemiology (Holmes et al., 2008; Viswan et al., 2019), genetics (Illig et al., 2009; Dehghan, 2019), organism classification (Bundy et al., 2002; Mahrous and Farag, 2015), and toxicology (Lindon et al., 2003; Hajduk et al., 2016). For instance, Bieleń et al. (2019) 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. These metabolic alterations are characteristic for malnutrition or cachexia and their early detection enables identifying and monitoring patients with a higher risk of weight loss.

One-dimensional  $^1\text{H}$  NMR spectroscopy (which we refer to as 1D NMR) remains a leading analytical technology in metabolomics, with advantages including high reproducibility, relatively rapid spectral acquisition times and NMR resonances that provide a direct measure of metabolite concentration based upon a single internal standard (Hore, 2015). Each metabolite in a 1D NMR spectrum presents a characteristic resonance, or peak, signature of intensity proportional to its concentration in the biological mixture. Typical biological mixtures often contain thousands of metabolites. Many of the resonance peaks generated by these metabolites create severe spectral overlaps, which seriously restricts the quantitative analysis of metabolites. Astle et al. (2012) developed a Bayesian model, which incorporates information available in online databases on the patterns of spectral resonances generated by human metabolites, to automate peak assignment and spectral deconvolution for 1D NMR spectra in the frequency domain. This model and its specially designed Markov chain Monte Carlo strategy are implemented in the R package BATMAN (Hao et al., 2012). However, this model cannot fully address the problem 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.

Two-dimensional (2D) NMR spectroscopy has considerable benefits over 1D NMR in metabolomics, as it can substantially improve spectral deconvolution and identification, at the expense of prolonged experimental time. Compared to 1D spectra, peak overlap in 2D spectra is greatly diminished because spin magnetization is transferred between different nuclear spins, resulting in more informative 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 the 2D  $^1\text{H}$  *J*-resolved NMR spectroscopy (JRES), first introduced by Aue et al. (1976b). *J*-coupling, also known as spin-spin coupling or scalar coupling, refers to the splitting of each resonance into multiple peaks,

due to the interaction between nearby nuclei. Specific chemical substructures (e.g. CH<sub>2</sub>-CH<sub>3</sub>, a carbon atom bonded to two hydrogen atoms and to another carbon atom which in turn is bonded to three hydrogen atoms) give rise to characteristic splitting patterns (e.g. doublets or triplets), leading each molecule to have a distinctive pattern of peaks which helps, for example, when identifying unknown molecules. *J*-coupling, moreover, has the advantage that the coupling patterns are less sensitive to changes in pH than chemical shift values (Moore and Sillerud, 1994). While 2D methods, such as correlation spectroscopy (COSY) (Aue et al., 1976a; Braunschweiler and Ernst, 1983) or total correlation spectroscopy (TOCSY) (Davis and Bax, 1985), use *J*-coupling to correlate chemical shifts of the coupling spins, JRES spectroscopy disperses the overlapping resonances into a second dimension and can provide a metabolic fingerprint in a relatively short acquisition time because of the low number of increments recorded in the indirect dimension. In 1D NMR 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 by using JRES spectroscopy therefore significantly reduces congestion, and enhances metabolite identification and estimation. For further details on JRES spectroscopy we refer to (Ludwig and Viant, 2010) and the references therein. The 2D JRES spectra are collections of convolved peaks, of which Figure 1 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. The 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, i.e. by the chemical structure of the molecules in which they are embedded and by the configuration of their chemical bonds within the molecules. Consequently, every metabolite has a characteristic molecular 2D JRES 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.

Standard analysis of JRES measurements is often based on 1D projections of the 2D spectra. For example, Viant (2003) perform multivariate statistical analyses for JRES metabolomics data by taking projections of each 2D spectrum onto the chemical shift axis. However, 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. This strategy is therefore not ideal as it does not allow the available information to be fully exploited. Gómez et al. (2014) 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. (2016) 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. (2007) 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 binned or original spectral data; and (iii) peak alignment followed by pattern recognition methods such as principal component analysis or partial least squares regression. The limitations of binning spectral data are well documented (Craig et al., 2006; Forgacs et al., 2011) 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 (Weljie et al., 2006), usually performed in 1D against a standard library of metabolite resonance peaks, reduces the requirement of expert spectroscopist knowledge but is still labour intensive.

**Contribution of this article:** 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 article, 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 known concentration.

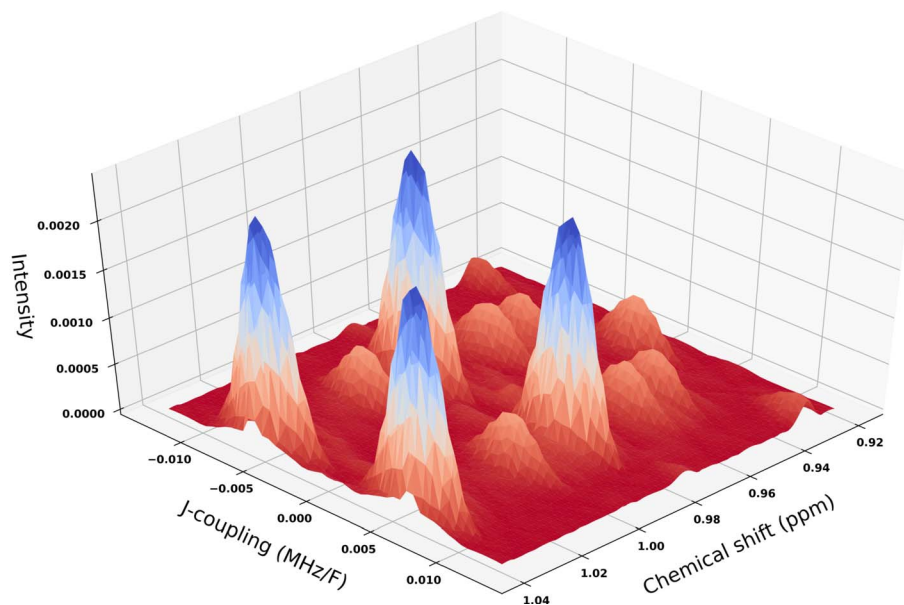


Figure 1: Example of a JRES spectrum surface plot. The  $x$ -axis corresponds to the chemical shift and is measured in parts per million (ppm) of the resonant frequency of a standard peak. The  $y$ -axis corresponds to the  $J$ -coupling information and shows the distance of each peak from the centre of the resonance measured in  $\text{Hz}/F$ . The standardized intensity on the  $z$ -axis is proportional to the concentration of the corresponding metabolite.

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 binning, 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.

## 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). 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 two main 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. After this standard preprocessing, which is typically performed fully automatically with the spectrometer manufacturer’s proprietary software (or using publicly available packages such as NMRglue (Helmus and Jaroniec, 2013)), a frequency-domain 2D JRES spectrum, as exemplified in Figure 1, is given by position vectors  $\mathbf{x} = (x_1, \dots, x_{N_C})$  on the chemical shift axis and  $\mathbf{y} = (y_1, \dots, y_{N_J})$  on the  $J$ -coupling axis, together with a measurement matrix  $\mathbf{z} = (z_{ij})_{i=1, \dots, N_C; j=1, \dots, N_J}$  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

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

The  $\phi$  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  $\xi$  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  $\phi$  component. This construction mirrors an equivalent modelling strategy developed by Astle et al. (2012) for 1D NMR data. We model the  $\phi$  component parametrically via continuous functions of continuous chemical shift and  $J$ -coupling information, using the physical theory of  $J$ -resolved NMR (see, e.g., Ludwig and Viant, 2010). The  $\xi$  component is modelled non-parametrically using a wavelet system constructed from a piecewise linear B-spline (see Dong and Shen, 2015).

## 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  $\delta$  on the chemical shift axis and the position  $\zeta$  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.

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

where the  $t_m$  are continuous template functions specifying the JRES signatures of the metabolites, with concentrations  $\beta_m$  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,

$$t_m(\delta, \zeta) = \sum_u \rho_{mu} g_{mu}(\delta - \delta_{mu}^*, \zeta), \quad (3)$$

where  $u$  is indexing the multiplets  $g_{mu}$  belonging to the  $m$ -th metabolite. The chemical shift parameter  $\delta_{mu}^*$  of the multiplet specifies the position of the centre of mass of  $g_{mu}$ . The coefficients  $\rho_{mu}$  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 (Hore, 2015) the  $\rho_{mu}$  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 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 2). 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 (5) below. More precisely,

$$g_{mu}(\delta, \zeta) = \sum_{v=1}^{V_{mu}} w_{muv} f_{\sigma_1 \sigma_2 \nu}(\delta, \zeta - \zeta_{muv}), \quad (4)$$

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  $\zeta_{muv}$  determine the  $J$ -coupling offsets of the peaks from the

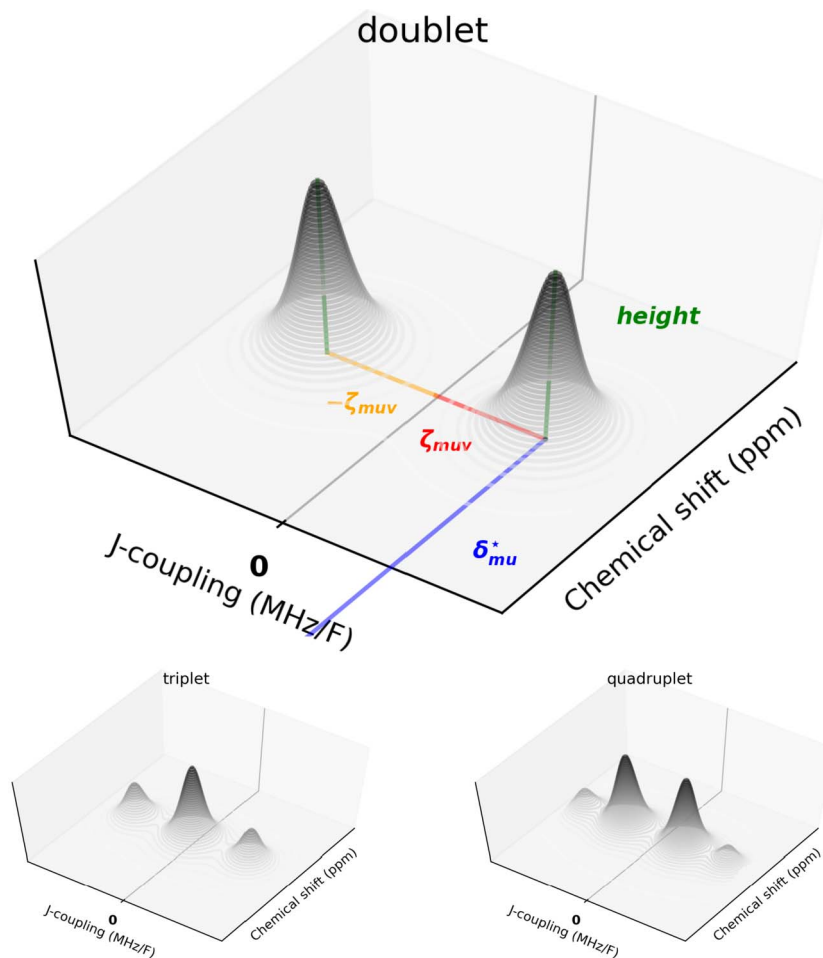


Figure 2: Peak configurations of some common multiplet types. The  $x$ -axis indicates chemical shift while the  $y$ -axis indicates the  $J$ -coupling. The upper panel shows a doublet with chemical shift  $\delta_{mu}^*$  and peak offset  $\zeta_{mu\nu}$ . The lower panel shows a triplet and quadruplet.

centre of mass of the multiplet. Multiplets are usually symmetric around  $\zeta = 0$ , with  $\{-\zeta_{mu\nu}\}_{\nu=1,\dots,V_{mu}} = \{\zeta_{mu\nu}\}_{\nu=1,\dots,V_{mu}}$ , and  $w_{mu\nu'} = w_{mu\nu}$  whenever  $\zeta_{mu\nu'} = -\zeta_{mu\nu}$ , see Figure 2.

Under ideal experimental conditions, the individual peaks in 1D NMR spectra have the shape of Lorentzians (Cavanagh et al., 2007). 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 (Goldman, 1992). In many types of spectroscopy, Voigt profiles are used to model peak



shapes (Bruce et al., 2000). 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

$$f_{\sigma_1\sigma_2\nu}(x, y) = \frac{\Gamma((\nu+2)/2)}{\Gamma(\nu/2)\pi\nu\sigma_1\sigma_2} \left(1 + \frac{1}{\nu} \left(\frac{x^2}{\sigma_1^2} + \frac{y^2}{\sigma_2^2}\right)\right)^{-(\nu+2)/2} \quad \text{for } (x, y) \in \mathbb{R}^2, \quad (5)$$

where  $\sigma_1, \sigma_2$  are scaling parameters controlling peak width,  $\nu$  represents the number of degrees of freedom controlling the tail decay, and  $\Gamma$  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 (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  $\nu$  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 Lorentzian 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  $\nu$  at large value, based on the observation that peaks in the data decay rapidly, and in general the choice of  $\nu$  should be dictated by the particular experimental conditions.

## 2.2 Modelling of uncatalogued metabolite signal

We model the uncatalogued component of (1) using a discrete B-spline wavelet tight frame. Frames have first been introduced by Duffin and Schaeffer (1952) and gained in popularity since the work of Daubechies et al. (1986). While frames are widely used in engineering applications (Mallat, 2008; Casazza and Kutyniok, 2012), they have been employed less in other fields. For a comprehensive introduction to wavelet frames we refer to Mallat (2008) and for further details on the particular systems described in this section to Dong and Shen (2010, 2015). 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 and high-pass details. 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 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. (2012) 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 a separable Hilbert space  $\mathcal{H}$  with inner product  $\langle \cdot, \cdot \rangle$  and a finite or countable index set  $I$ , a sequence  $\{g_i\}_{i \in I} \subset \mathcal{H}$  is called a *tight frame* for  $\mathcal{H}$  if

$$f = \sum_{i \in I} \langle f, g_i \rangle g_i \quad \text{for all } f \in \mathcal{H}. \quad (6)$$

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 (6) to be unique. Indeed, the only properties of Hilbert space orthonormal bases that (Astle et al., 2012) use for their inferential method is (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  $f \in \mathcal{H}$  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  $\sum_{i \in I} c(i)g_i \in \mathcal{H}$ . 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  $\mathcal{H}$ .

The elements of a wavelet frame are generated by shifts and dilations of, in general more than one, generators, called *framelets*. In this article, we use a *discrete B-spline wavelet tight frame*. This class of frames is widely used in wavelet frame based image restoration and has first been introduced by Ron and Shen (1997). 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

$$\{(\mathbf{a}^{(l)}(n - k))_{n \in \mathbb{Z}^d} : l \in \{1, \dots, r\}, k \in \mathbb{Z}^d\}, \quad (7)$$

consisting of all of their integer shifts. Sufficient for the system (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 (1997), in which case the analysis and synthesis operators are given via discrete convolutions by

$$\mathbf{W}: \mathbf{u} \in \ell_2(\mathbb{Z}^d) \rightarrow \left( \sum_{j \in \mathbb{Z}^d} \mathbf{a}^{(l)}(j - k) \mathbf{u}(j) \right)_{(k,l) \in \mathbb{Z}^d \times \{1, \dots, r\}} \in \ell_2(\mathbb{Z}^d \times \{1, \dots, r\}) \quad (8)$$

and

$$\mathbf{W}^\top: c \in \ell_2(\mathbb{Z}^d \times \{1, \dots, r\}) \rightarrow \left( \sum_{l=1}^r \sum_{j \in \mathbb{Z}^d} c(k - j, l) \mathbf{a}^{(l)}(j) \right)_{k \in \mathbb{Z}^d} \in \ell_2(\mathbb{Z}^d). \quad (9)$$

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 (8) and (9) describe the undecimated single level fast wavelet transform. 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 (8) and (9) are performed using symmetric boundary extensions matching the symmetry of the respective filters. 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  $\max(1 - |x|, 0)$ , while the highpass filters  $\mathbf{a}^{(2)}$ , respectively  $\mathbf{a}^{(3)}$ , are wavelet masks of piecewise linear anti-symmetric, respectively symmetric, framelets. 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  $3 \times 3$ .

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 (Mallat, 2008), 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.

### 2.3 Likelihood

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

$$\mathbf{Wz} = \sum_{m=1}^M \beta_m \mathbf{W}\mathbf{T}_m + \boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad \epsilon_{ij\ell} \sim \text{N}(0, \lambda^{-1}), \quad (10)$$

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  $\boldsymbol{\epsilon} = (\epsilon_{ij\ell}) \in \mathbb{R}^{N_C \times N_J \times r}$  are independent identically Normal distributed errors with scalar precision parameter  $\lambda$ . For every  $l = 1, \dots, r$ , 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  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_M)^\top$  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. (2012) by imposing a hard thresholding constraint in signal domain, their approach makes computations inefficient and therefore infeasible for the 2D setting. In Sections 3 and 6, we compare our approach with the prior and wavelet specifications of Astle et al. (2012) and highlight the advantages of our method.

## 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 (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. (2012) 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 Astle et al., 2012). 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 distribution 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  $\theta$  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 (Mitchell and Beauchamp, 1988; George and McCulloch, 1993) and a variety of continuous shrinkage priors (see, for example, Polson et al., 2012; Bhattacharya et al., 2015; Piironen et al., 2017; Bhadra et al., 2019). The spike-and-slab prior is intuitively appealing as it performs 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 (Carvalho et al., 2010), the LASSO (Tibshirani, 1996) and the Student- $t$  prior (Tipping, 2001). We use the horseshoe prior since its flat Cauchy-like tails allow components of  $\theta$  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  $\theta$ . We further make use of the localization of the framelets to additionally shrink the framelet coefficients  $\theta$  in regions of targeted metabolites.

In more detail, given a global shrinkage parameter  $\tau$ , the horseshoe prior for  $\theta_{ijl}$  can be represented as the scaled mixture of Normals

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

where the  $\theta_{ijl} \mid \tau$  are conditionally independent and where the local shrinkage parameters  $\mu_{ijl}$  are assigned half Cauchy distributions. As suggested by Gelman (2006), we also assign a half Cauchy distribution to the *global* shrinkage parameter,  $\tau \sim C^+(0, d)$ . 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. 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 8 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  $\lambda$ . 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  $\delta_{mu}^*$  and  $\zeta_{mu\nu}$  in frequency domain (see Figure 2). Using prior information about the uncertainty of these parameters, discussed below, we determine regions, centered around  $(\delta_{mu}^*, \zeta_{mu\nu})$ , of likely locations for the specified multiplets and identify the index set  $\mathcal{I} \times \mathcal{J} \subset N_C \times N_J$  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  $(i, j) \in \mathcal{I} \times \mathcal{J}$  and  $\omega_{ij} = c_l$  if  $(N_C \times N_J) \setminus (\mathcal{I} \times \mathcal{J})$ . Next, define the hyperparameters  $c_{ijl}$  controlling the local shrinkage of the coefficients  $\theta_{ijl}$  of the  $l$ -th framelet filter ( $l = 1, \dots, r$ ) 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  $(\omega_{ij})$  in signal domain. Specifically, noting that all filters we use have support of size  $3 \times 3$ , 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 3 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 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 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 multiplet. 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

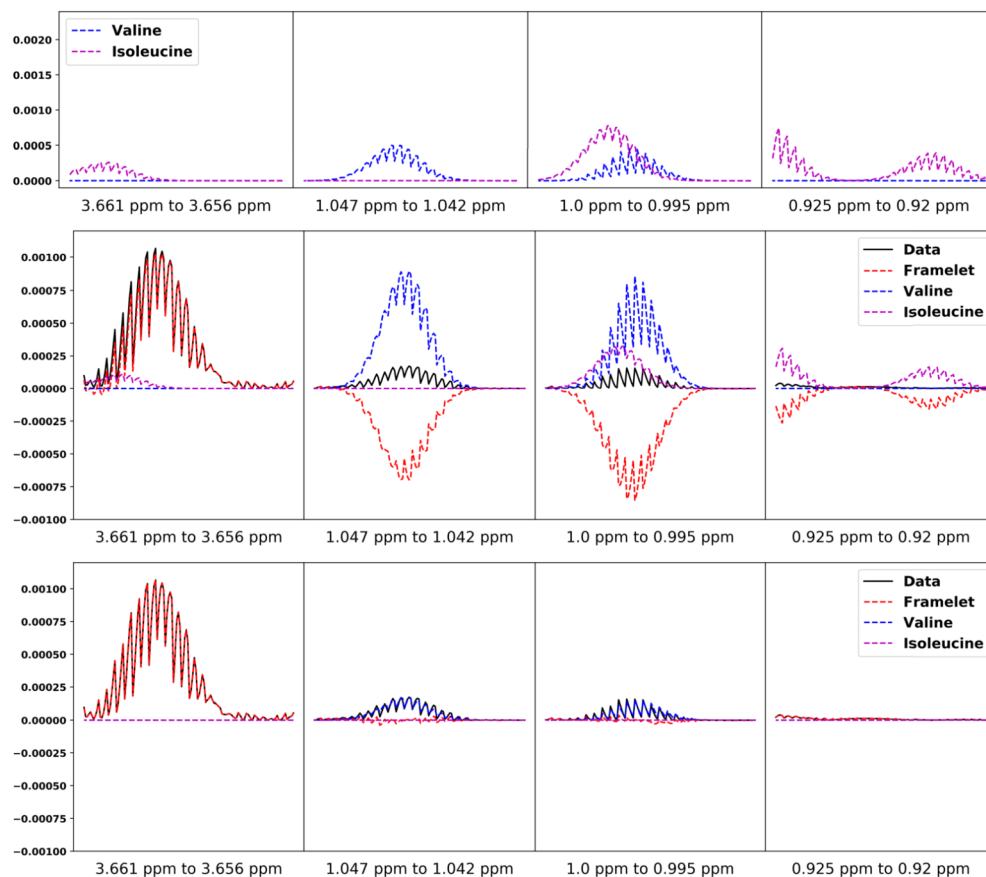


Figure 3: Effect of additional local shrinkage applied to framelet coefficients of selected targeted regions. For ease of visualization, spectra are vectorised columnwise and plotted in 2D. On the  $x$ -axis we report the chemical shift region of the multiplet, and on the  $y$ -axis their intensities. 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.

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 multiplet at 1.045ppm is driving the estimation of concentration, 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  $\lambda$ . 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 and the region around 1.045ppm is driving the estimation of concentration of Valine. These regions are the least overlapped for the two metabolites. 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. (2012) to our 2D model.

*Prior for precision parameter  $\lambda$ :* 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  $\lambda$ . For the simulations and 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 spectra, their changes are negligible within spectra. We therefore assume that peaks within a spectrum are dependent upon two global peak width parameters  $\sigma_1$  and  $\sigma_2$ , see (5), for which we choose log-Normal distributions with median  $1\text{Hz}/F$  and variance  $4.6\text{Hz}^2/F^2$ , where  $F$  is the operating frequency of the spectrometer in MHz. These priors give good support to a broad region around  $1\text{Hz}/F$ , the typical peak widths generated by modern spectrometers (Hore, 2015). 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  $\nu$ , see (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 6, we prefer to fix  $\nu$ .



*Priors for multiplets:* The parametrization of metabolite signature templates is done in two steps, see (3) and (4), via linear combinations of multiplets along the chemical shift axis, which in turn arise as linear combinations of Student- $t$  distributions (5) along the  $J$ -coupling axis. Uncertainty of peak positions can therefore be modelled separately within and between multiplets. The parameters  $\zeta_{muv}$  and  $w_{muv}$ , determining the peak positions on the  $J$ -coupling axis and their amplitudes within multiplets in (4), can be computed via simple rules from the  $J$ -coupling constants  $J_{mu}$  (see Hore (2015) for details) and may be assumed to be constant across spectra. The multiplet chemical shift parameters  $\delta_{mu}^*$  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  $\hat{\delta}_{mu}^*$  for  $\delta_{mu}^*$  are published in online databases and can be used to construct an informative prior distribution. The deviations of both  $J_{mu}$  and  $\delta_{mu}^*$  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  $7\text{Hz}^2$ , and truncation region  $[\frac{1}{2}\hat{J}_{mu}, \frac{3}{2}\hat{J}_{mu}]$ . For each  $\delta_{mu}^*$  we choose a truncated Normal prior distribution with mean  $\hat{\delta}_{mu}^*$ , variance  $10^{-4}\text{ppm}$ , and truncation region  $[\hat{\delta}_{mu}^* - 0.03\text{ppm}, \hat{\delta}_{mu}^* + 0.03\text{ppm}]$ . 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  $\delta_{mu}^*$ .

*Priors for metabolite abundances:* Each coefficient  $\beta_m$  in (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  $\beta_m$  is restricted to  $[0, \infty)$ . Conjugacy considerations motivate the use of a truncated Normal prior distribution for each component, i.e.  $\beta_m \sim \text{TN}(e_m, 1/s_m^2, 0, \infty)$ . 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  $s_m^2 = 10^{-6}$  for all  $m = 1, \dots, M$ , indicating low prior information.

## 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. (2012), 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 (Heinecke et al., 2020).

We employ Gibbs samplers to update the components of  $\beta$  and  $\theta$ , both having truncated Normal conditional distributions, and the precision parameter  $\lambda$ , 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  $\sigma_1$  and  $\sigma_2$  we use log-Normal proposals. To update the multiplet chemical shift parameter  $\delta_{mu}^*$ , we propose  $\delta_{mu}^{*'}$  from the truncated Normal distribution

$$\text{TN}\left(\delta_{mu}^*, V_{\delta_{mu}^*}^2, \delta_{mu}^* - 0.03\text{ppm}, \delta_{mu}^* + 0.03\text{ppm}\right)$$

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

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

For the local shrinkage parameters  $\mu_{ijl}$  and the global shrinkage parameter  $\tau$  we employ Gaussian proposals truncated below at zero. All proposal variances are adapted using the adaptive Metropolis-within-Gibbs algorithm of Roberts and Rosenthal (2009), 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 not utilised in the examples of this article.

## 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, Ulrich et al., 2007). 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  $i$ th simulated biological mixture is

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

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  $\epsilon$  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

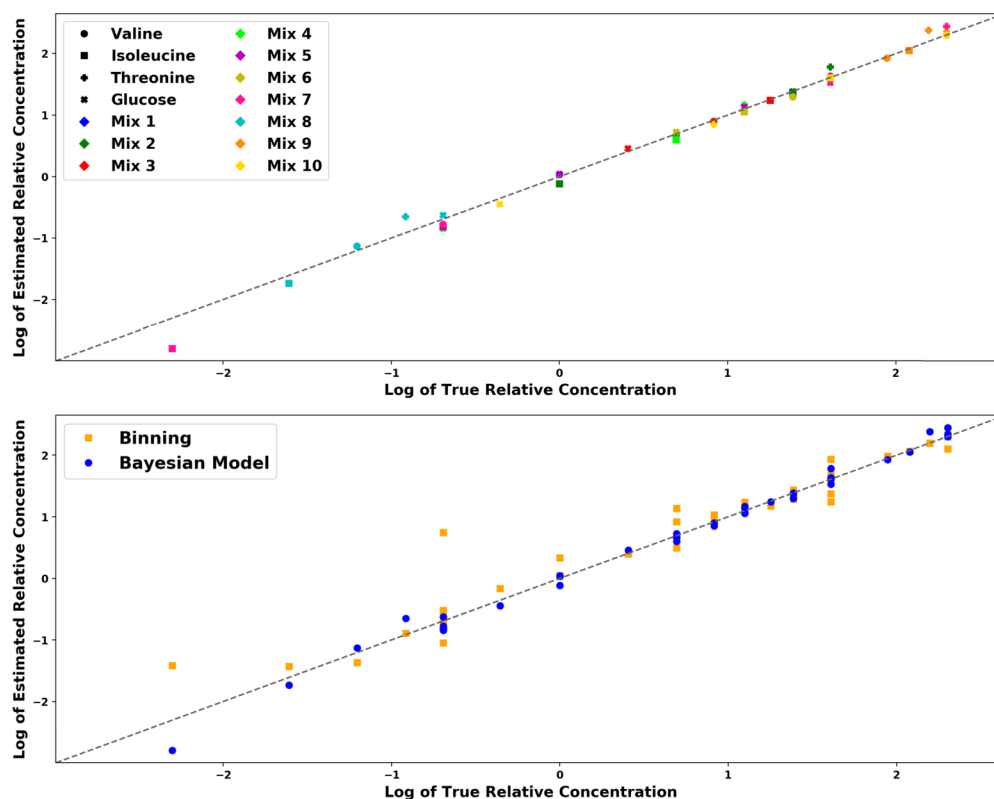


Figure 4: Top panel: Comparison between the logarithm of the true relative concentrations and the estimated relative concentrations obtained with our method on the ten mixtures. Bottom panel: Performance comparison between our approach and the binning on the ten simulated biological mixtures.

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. (2010), for the choice of  $d$  and  $c_h$  see Section 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 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 binning the spectral data, which is commonly done in metabolic analysis (see, for example, Sousa et al., 2013). 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 it is clear that binning does not perform as well. Further details on the simulation results and the comparison are presented in Supplementary Material S2.

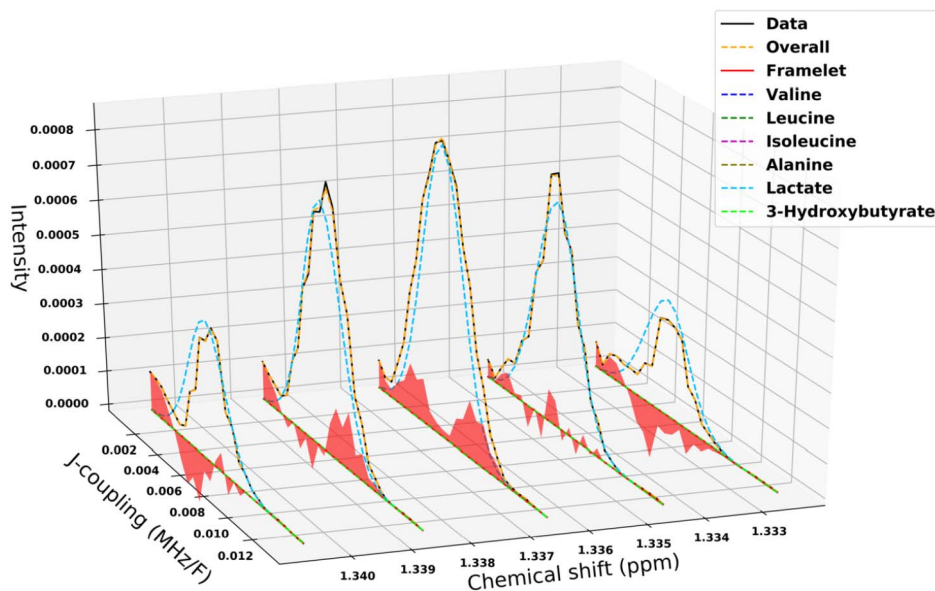


Figure 5: Deconvolution surface plot from urine JRES spectrum for the region around 1.337ppm, where the resonance is generated by Lactate. For ease of visualization we plot the fit on a grid of equally spaced points with distance 0.002ppm for the chemical shift axis.

## 6 Performance on urine and serum spectra

We examine the performance of our method on a urine and a serum dataset.  $^1\text{H}$  NMR spectra of human urine and serum samples were obtained from healthy participants of the Airwave Health Monitoring Study (Elliott et al., 2014). The samples were prepared and acquired according to protocols published in Dona et al. (2014). 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 jresgpprpf 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 the singlet signal of TSP set at 0ppm (urine) or to the doublet resonance of  $\alpha$ -glucose set at 5.23ppm (serum) using Topspin 3.2 software (Bruker Biospin Ltd).

## 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-Hydroxybutyrate. A second performance demonstration on the 2D JRES human serum spectrum is included in Section S4 of the Supplementary Material, and yields results broadly similar to the urine spectrum. A sensitivity analysis on the 2D JRES human urine spectrum is included in Section S5 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 5 shows a surface plot of the metabolite fit around 1.337ppm, while Figure 6 shows heat maps of the measured spectrum, overall fitting and metabolite fitting. Along with the additional column-wise 2D plots of the metabolite estimations provided in Figures S3 and S4 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  $7.721 \times 10^{-9}$ . For Valine (Figure S3, 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 S3, 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 S4, 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 S4, 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 S5–S10 in Supplementary Material show traceplots of the log-likelihood, of some randomly selected framelet coefficients, of the precision parameter  $\lambda$ , of the concentration  $\beta$ , of the chemical shift  $\delta$  and of the  $J$ -coupling  $\zeta$  parameters for selected metabolites. We report the results obtained from running three different chains. It is worth noticing that the dimension of the parameter space is greater than 204000. While it can be seen that the framelet coefficients and the

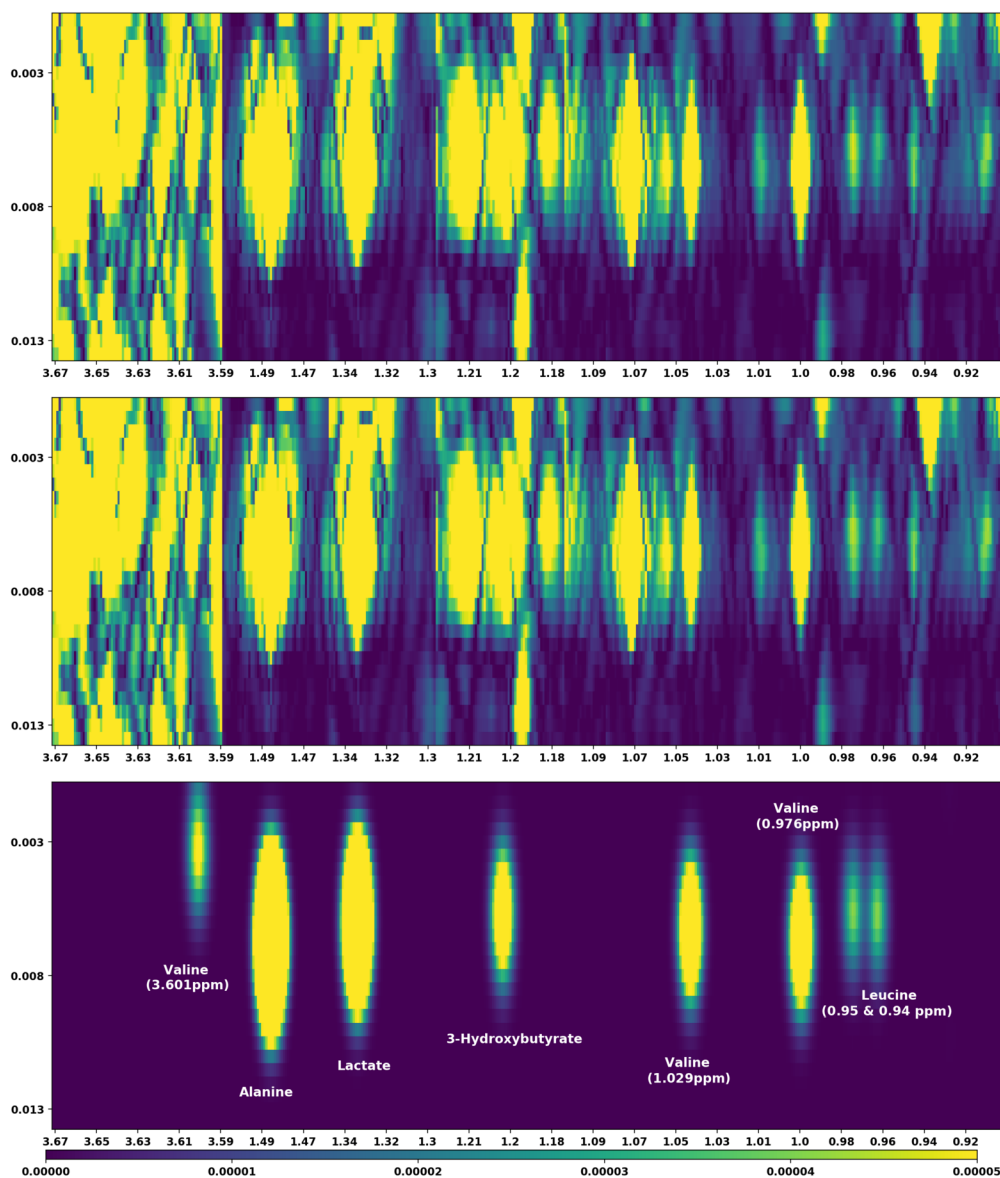


Figure 6: Heat maps for intensities from the urine JRES dataset. The  $x$ -axis corresponds to chemical shift in ppm, the  $y$ -axis to  $J$ -coupling in MHz/F. Plots show original data (upper panel), overall fitting, i.e., metabolite and framelet fitting (middle panel), and fitting of metabolites only (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 estimate is close to zero.

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 S5 in Supplementary Material it can be seen that the traceplot of the log-likelihood is satisfactory (Robert and Casella, 2013). Further, unidentifiability issues often lead to the presence of ridges in the posterior, as in our case. For the parameters defining the catalogued metabolites (concentration,  $J$ -coupling and chemical shift parameters), which are those most affected by the identifiability problems, the traceplots show that the three chains end in different – but very close – regions of the posterior parameter space. Nevertheless, the traceplots of the log-likelihood (where the three chains overlap) show that the algorithm finds a region of high likelihood, as it seems to be able to reach the ridge in the posterior and moves within it. This is confirmed also by Table S3 in Supplementary Material, which compares the estimates of the relative concentrations obtained from different chains. These estimates are consistent, except for Isoleucine which is present in the sample at almost zero concentration, and as such it is more affected by structural noise.

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.

## 6.2 Methods comparison

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 overlap (Fonville et al. (2010); Féraud et al. (2015)). 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 a baseline metabolite we choose Valine, since it is relatively isolated in both the 1D and the 2D spectra. In Figure 7 we compare the estimation results for relative concentrations obtained via our method applied to 1D NMR data, 2D JRES and via binning. (For the numerical values see Table S2 in Supplementary Material.) Note that binning 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 data do not overlap, clearly showing the potential of 2D NMR spectroscopy. Note that Figure 8 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

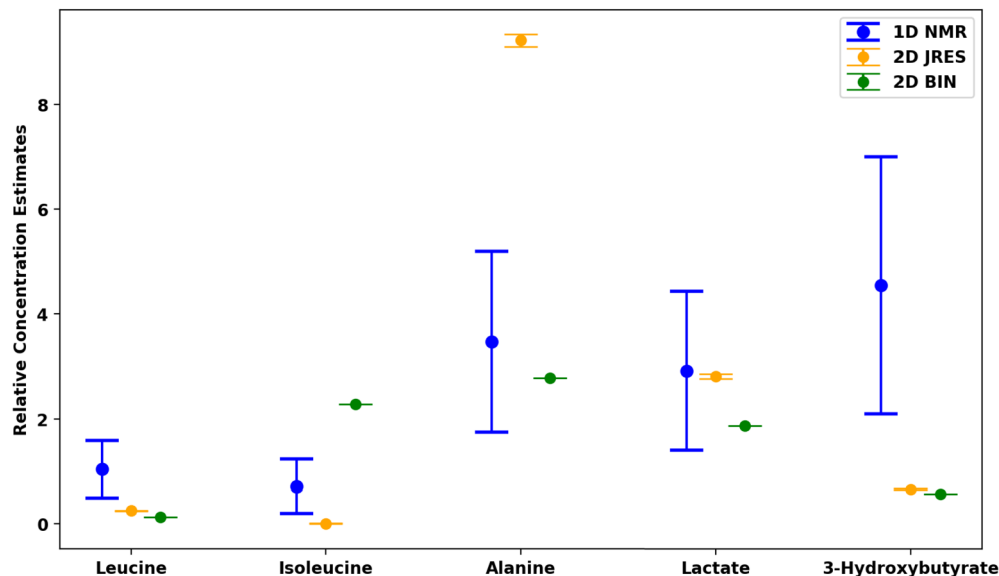


Figure 7: Posterior relative concentration estimates and posterior standard deviations using our method on the urine spectrum from 1D NMR measurements, from 2D JRES measurements as well as using binning on the 2D JRES measurements. Valine is chosen as baseline. For four of the five targeted metabolites the posterior means of the estimates obtained using the second dimension differ by more than 25%. The figure shows 95% credible intervals. Note binning only produces point estimates with no quantification of uncertainty of the estimate.

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 6. 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 8 (around 1.49ppm). While this problem could be overcome by allowing the distance between peaks in a multiplet to fluctuate around the theoretical value, we do not deem the introduction of the additional computational burden necessary as the problem rarely occurs on a spectrum. Moreover, when dealing with urine, a further obstacle to identification 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 6, 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 S2 in Supplementary Material. Traditional binning has



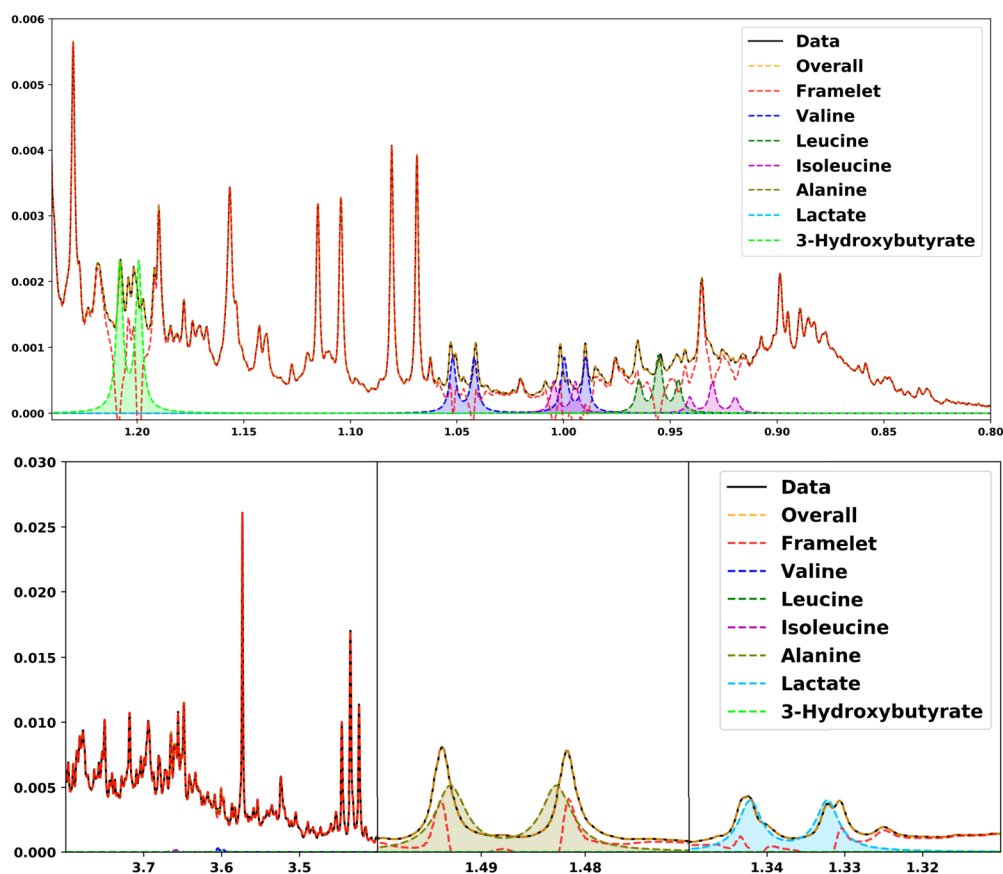


Figure 8: Deconvolution of selected regions from the urine 1D NMR data. The  $x$ -axis corresponds chemical shift in ppm and  $y$ -axis to intensities. The top panel shows resonances generated by Valine, Leucine, Isoleucine and 3-Hydroxybutyrate. The lower middle panel and lower right panel show resonances generated by Alanine and Lactate, respectively. The lower left panel shows resonances generated by untargeted metabolites and weak signals from Valine and Isoleucine.

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.

### 6.3 1D NMR spectra and comparison with BATMAN

The R package BATMAN (Bayesian automated analyzer for NMR, see Hao et al., 2012, 2014) implements the Bayesian method for 1D NMR introduced by Astle et al. (2012), but currently cannot be run on 2D NMR data. We therefore compare our method with

	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 1: Comparison of modelling strategy between BATMAN and our approach.

BATMAN on the 1D human urine data set. Notice that our approach is also suitable to analyse 1D NMR spectra (see Section 6.2 below), as it improves on the original strategy adopted in BATMAN. The main modelling differences between our work and the paper by Astle et al. (2012) are summarised in Table 1. Our improvements have led to a more 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. (2012). Moreover, theoretically,  $J$ -coupling constants vary only insignificantly between spectra, motivating Astle et al. (2012) 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 8 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  $1.195 \times 10^{-5}$  for our method and  $1.193 \times 10^{-5}$  for BATMAN, which shows a good performance of both methods. Nevertheless the main limitation 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 2 shows a comparison between the summary statistics of the effective sample sizes (ESS) (Ripley, 2009) and of the integrated autocorrelation times (IAC) (Christen and Fox, 2010; Kalli et al., 2011) 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

		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 2: 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.

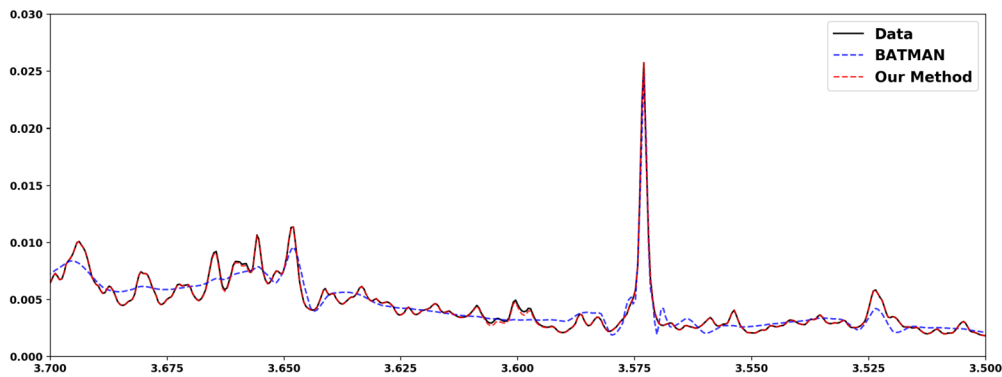


Figure 9: Deconvolution of resonances generated by untargeted metabolites for a selected region from a urine 1D NMR spectrum. The  $x$ -axis corresponds to chemical shift in ppm and the  $y$ -axis to the intensities. The measured spectrum is shown in black, while the B-spline frame component of our model is plotted in red and the Symlet 6 wavelet component of BATMAN in blue.

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 supported by comparing the IACs: once again, on average, the posterior estimation from our method is more accurate and mixing is improved. Figure 9 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.

## 7 Discussion

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.

## Supplementary Material

Supplementary Materials (DOI: [10.1214/20-BA1208SUPP](https://doi.org/10.1214/20-BA1208SUPP); .pdf).

## References

- 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. MR3036393. doi: <https://doi.org/10.1080/01621459.2012.695661>. 426, 430, 434, 436, 440, 441, 449, 450

- 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. doi: <https://doi.org/10.1063/1.432450>. 427
- 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. doi: <https://doi.org/10.1063/1.431994>. 426
- Bhadra, A., Datta, J., Polson, N. G., Willard, B., et al. (2019). “Lasso meets horseshoe: A survey.” *Statistical Science*, 34(3): 405–427. MR4017521. doi: <https://doi.org/10.1214/19-STS700>. 437
- Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). “Dirichlet–Laplace priors for optimal shrinkage.” *Journal of the American Statistical Association*, 110(512): 1479–1490. MR3449048. doi: <https://doi.org/10.1080/01621459.2014.960967>. 437
- 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. 426
- Braunschweiler, L. and Ernst, R. (1983). “Coherence transfer by isotropic mixing: Application to proton correlation spectroscopy.” *Journal of Magnetic Resonance*, 53(3): 521–528. URL <http://www.sciencedirect.com/science/article/pii/0022236483902263>. 427
- 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  $^1\text{H}$ -NMR-based metabolomics.” *Nature Medicine*, 8: 1439–1445. doi: <https://doi.org/10.1038/nm1202-802>. 426
- 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. 433
- 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. URL <http://www.sciencedirect.com/science/article/pii/S0014579302028545>. 426
- Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). “The horseshoe estimator for sparse signals.” *Biometrika*, 97(2): 465–480. URL <http://www.jstor.org/stable/25734098>. MR2650751. doi: <https://doi.org/10.1093/biomet/asq017>. 437, 443
- Casazza, P. G. and Kutyniok, G. (2012). *Finite Frames: Theory and Applications*. Birkhäuser Basel. MR2964005. doi: <https://doi.org/10.1007/978-0-8176-8373-3>. 433

- Cavanagh, J., Skelton, N., Fairbrother, W., Rance, M., Palmer, A., Skelton, N., and Rance, M. (2007). *Protein NMR Spectroscopy: Principles and Practice*. Academic Press. 432
- Christen, J. A. and Fox, C. (2010). “A general purpose sampling algorithm for continuous distributions (the *t*-walk).” *Bayesian Analysis*, 5(2): 263–281. MR2719653. doi: <https://doi.org/10.1214/10-BA60>. 450
- 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): 2262–2267. 428
- Daubechies, I., Grossmann, A., and Meyer, Y. (1986). “Painless Nonorthogonal Expansions.” *Journal of Mathematical Physics*, 27. MR0836025. doi: <https://doi.org/10.1063/1.527388>. 433
- 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. doi: <https://doi.org/10.1021/ja00295a052>. 427
- Dehghan, A. (2019). “Linking Metabolic Phenotyping and Genomic Information.” In *The Handbook of Metabolic Phenotyping*, 561–569. Elsevier. 426
- 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. doi: <https://doi.org/10.1021/ac5025039>. 444
- Dong, B. and Shen, Z. (2010). “MRA-based wavelet frames and applications.” *IAS/Park City Mathematics Series*, 19. MR3098080. 433
- Dong, B. and Shen, Z. (2015). “Image restoration: a data-driven perspective.” In *Proceedings of the International Congress on Industrial and Applied Mathematics (ICIAM)*, 65–108. Beijing, China: High Education Press. MR3408465. 430, 433
- Duffin, R. and Schaeffer, A. (1952). “A class of nonharmonic Fourier series.” *Transactions of the American Mathematical Society*, 72: 341–366. MR0047179. doi: <https://doi.org/10.2307/1990760>. 433
- 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. URL <http://www.sciencedirect.com/science/article/pii/S0013935114002564>. 444
- 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. URL <http://libproxy1.nus.edu.sg/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=110339434&site=ehost-live>. 447

- Fonville, J. M., Maher, A. D., Coen, M., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2010). "Evaluation of Full-Resolution J-Resolved  $^1\text{H}$  NMR Projections of Biofluids for Metabonomics Information Retrieval and Biomarker Identification." *Analytical Chemistry*, 82(5): 1811–1821. PMID: 20131799. doi: <https://doi.org/10.1021/ac902443k>. 447
- 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. 428
- Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1(3): 515–534. MR2221284. doi: <https://doi.org/10.1214/06-BA117A>. 437
- George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 437
- Goldman, M. (1992). *Quantum description of high-resolution NMR in liquids*. Oxford University Press. 432
- 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  $^1\text{H}$ -NMR data." *Analytical and Bioanalytical Chemistry*, 406(30): 7967–7976. doi: <https://doi.org/10.1007/s00216-014-8225-6>. 427
- 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. 426
- Hajduk, A., Mrochem-Kwarciak, J., Skorupa, A., Ciszek, M., Heyda, A., Skłodowski, K., Sokół, M., et al. (2016). " $^1\text{H}$  NMR based metabolomic approach to monitoring of the head and neck cancer treatment toxicity." *Metabolomics*, 12(6): 102. 426
- 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. doi: <http://dx.doi.org/10.1093/bioinformatics/bts308>. 426, 449
- Hao, J., Liebeke, M., Astle, W., De Iorio, M., Bundy, J., and Ebbels, T. M. D. (2014). "Bayesina deconvolution and quantification of metabolites in complex 1D NMR spectra using Batman." *Nature Protocols*, 9(6): 1416. 449
- Heinecke, A., Ye, L., De Iorio, M., and Ebbels, T. (2020). "Supplementary Materials." *Bayesian Analysis*. doi: <https://doi.org/10.1214/20-BA1208SUPP>. 441
- 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. 430

- Hollinshead, K. E., Williams, D. S., Tennant, D. A., and Ludwig, C. (2016). “Probing cancer cell metabolism using NMR spectroscopy.” In *Tumor Microenvironment*, 89–111. Springer. 426
- 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. (2008). “Human metabolic phenotype diversity and its association with diet and blood pressure.” *Nature*, 453: 396–400. doi: <https://doi.org/10.1038/nature06882>. 426
- Hore, P. (2015). *Nuclear Magnetic Resonance*. Oxford chemistry primers. Oxford University Press. URL <https://books.google.com.sg/books?id=L9umCAAQBAJ>. 426, 431, 440, 441
- 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: 137–141. doi: <https://doi.org/10.1038/ng.507>. 426
- Kalli, M., Griffin, J. E., and Walker, S. G. (2011). “Slice sampling mixture models.” *Statistics and Computing*, 21: 93–105. MR2746606. doi: <https://doi.org/10.1007/s11222-009-9150-y>. 450
- 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. doi: <https://doi.org/10.1021/acs.analchem.5b02311>. 427
- 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. URL <http://www.sciencedirect.com/science/article/pii/S0041008X02000790>. MR2077534. doi: [https://doi.org/10.1142/9789812388759\\_0013](https://doi.org/10.1142/9789812388759_0013). 426
- 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. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/pca.1186>. 427, 430
- 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. 426



- Mallat, S. (2008). *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, Inc. MR2479996. 433, 435
- Mitchell, T. J. and Beauchamp, J. J. (1988). “Bayesian variable selection in linear regression.” *Journal of the American Statistical Association*, 83(404): 1023–1032. MR0997578. 437
- 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. 427
- 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. 426
- 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. 428
- Piironen, J., Vehtari, A., et al. (2017). “Sparsity information and regularization in the horseshoe and other shrinkage priors.” *Electronic Journal of Statistics*, 11(2): 5018–5051. MR3738204. doi: <https://doi.org/10.1214/17-EJS1337SI>. 437
- 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. MR3000018. doi: <https://doi.org/10.1214/12-BA730>. 437
- 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: 45–50. doi: <https://doi.org/10.1038/83496>. 426
- Ripley, B. (2009). *Stochastic Simulation*. Wiley Series in Probability and Statistics. Wiley. URL <https://books.google.co.uk/books?id=rmGfsJxRDqgC>. MR2299137. 450
- Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media. MR1707311. doi: <https://doi.org/10.1007/978-1-4757-3071-5>. 447
- Roberts, G. O. and Rosenthal, J. S. (2009). “Examples of Adaptive MCMC.” *Journal of Computational and Graphical Statistics*, 18(2): 349–367. MR2749836. doi: <https://doi.org/10.1198/jcgs.2009.06134>. 442
- 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. MR1469348. doi: <https://doi.org/10.1006/jfan.1996.3079>. 434, 435
- 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. 443

- Tibshirani, R. (1996). “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1): 267–288. URL <http://www.jstor.org/stable/2346178>. MR1379242. 437
- Tipping, M. E. (2001). “Sparse Bayesian Learning and the Relevance Vector Machine.” *Journal of Machine Learning Research*, 1: 211–244. MR1875838. doi: <https://doi.org/10.1162/15324430152748236>. 437
- 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. doi: <https://doi.org/10.1093/nar/gkm957>. 442
- Viant, M. R. (2003). “Improved methods for the acquisition and interpretation of NMR metabolomic data.” *Biochemical and Biophysical Research Communications*, 310(3): 943–948. URL <http://www.sciencedirect.com/science/article/pii/S0006291X03018618>. 427
- 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*. doi: <https://doi.org/10.1002/nbm.4192>. 426
- Weljie, A. M., Newton, J., Mercier, P., Carlson, E., and Slupsky, C. M. (2006). “Targeted Profiling: Quantitative Analysis of <sup>1</sup>H NMR Metabolomics Data.” *Analytical Chemistry*, 78(13): 4430–4442. doi: <https://doi.org/10.1021/ac060209g>. 428

#### Acknowledgments

The authors thank Dr Goncalo Miguel Gomes Da Graca (Department of Surgery & Cancer, Imperial College London) for providing the urine and serum datasets.