A Bayesian Model of NMR Spectra for the Deconvolution and Quantification of Metabolites in Complex Biological Mixtures

*William Astle, *Maria De Iorio, *Sylvia Richardson,

†David Stephens and ‡Timothy Ebbels

*Department of Epidemiology and Biostatistics and

‡Section of Biomolecular Medicine, Department of Surgery and Cancer

Faculty of Medicine,

Imperial College, London.

United Kingdom.

†Department of Mathematics and Statistics,

McGill University, Montreal.

Canada.

May 12, 2011

Abstract

Nuclear Magnetic Resonance (NMR) spectra are widely used in metabolomics to obtain profiles of metabolites dissolved in biofluids such as cell supernatants. Methods for estimating metabolite concentrations from these spectra are presently confined to manual peak fitting and to binning procedures for integrating resonance peaks. Extensive information on the patterns of spectral resonance generated by human metabolites is now available in online databases. By incorporating this information into a Bayesian model we can deconvolve resonance peaks from a spectrum and obtain explicit concentration estimates for the corresponding metabolites. Spectral resonances that cannot be deconvolved in this way may also be of scientific interest so we model them jointly using wavelets.

We describe a Markov chain Monte Carlo algorithm which allows us to sample from the joint posterior distribution of the model parameters, using specifically designed block updates to improve mixing. The strong prior on resonance patterns allows the algorithm to identify peaks corresponding to particular metabolites automatically, eliminating the need for manual peak assignment. We assess our method for peak alignment and concentration estimation. Except sometimes when the target resonance signal is very weak, alignment is unbiased and precise. We compare the Bayesian concentration estimates to those obtained from a conventional numerical integration method and find that our estimates have 11 fold reduced mean estimation error.

Finally, we apply our method to a spectral dataset taken from an investigation of the metabolic response of yeast to recombinant protein expression. We compare our results to the manual deconvolutions of an expert spectroscopist and find concurrence except in the case of one compound. We discuss the reason for such discrepancies and the robustness of our methods concentration estimates.

KEYWORDS: metabolomics, concentration estimation, prior information, multi component model, block updates.

1. INTRODUCTION

Metabolomics (also known as *metabonomics* or sometimes *metabolic profiling*) is a scientific discipline concerned with the quantitative study of metabolites, the small molecules that participate in metabolic reactions. Research in this field is expanding rapidly, with applications in many areas of biology and medicine including cancer (e.g. Griffiths et al. (2002)), toxicology (e.g. Lindon et al. (2003)), organism classification (e.g. Bundy et al. (2002)), genetics (e.g. Illig et al. (2010)), biochemistry (e.g. Raamsdonk et al. (2001)), epidemiology (e.g. Holmes et al. (2008)) and disease diagnostics (e.g Brindle et al. (2002)).

Almost all experiments in metabolomics rely on measurements of the abundances of metabolites in complex biological mixtures, often biofluid or tissue samples. One of the most extensively used techniques for obtaining such quantitative information is proton nuclear magnetic resonance (^{1}H NMR) spectroscopy. Metabolites generate characteristic resonance signatures in ^{1}H NMR spectra and each signature appears with intensity proportional to the concentration of the corresponding metabolite in the biological mixture.

Specialized models and tools are needed to draw inferences from ^{1}H NMR spectroscopic datasets, which are large and heavily structured. At present there is no statistical method for analyzing metabolomic NMR spectra reflecting the data generating mechanisms and the extensive prior knowledge available, e.g. on the form of metabolite NMR signatures. In this paper, we describe novel Bayesian approaches for the analysis of ^{1}H NMR data from complex biological mixtures. Specifically, we develop new models reflecting the data generation mechanisms and our prior knowledge, using a combination of parametric functions and wavelets. We introduce a computational strategy based on Markov chain Monte Carlo (MCMC), including novel block updates to overcome strong posterior correlation between the parametric functions

and the wavelets. We demonstrate the utility of the approach with simulations and analyses of data from a yeast metabolomics experiment.

1.1 NMR spectroscopy

An NMR spectrum consists of a series of measurements of resonance intensity usually taken on a grid of equally spaced frequencies. Figure 1 is a representative section (of about one tenth) of an NMR spectrum taken from an experiment into the metabolomic response of yeast to recombinant protein expression. The x-axis of the spectrum corresponds to resonant frequency and the y-axis to resonance intensity.

[Figure 1 about here.]

The spectrum is a collection of convolved peaks with different horizontal positions and vertical scalings, each of which has the form of a Lorentzian curve. The zero centered, standardized Lorentzian has equation

$$l_\gamma(x) = \frac{2}{\pi}\frac{\gamma}{4x^2+\gamma^2}. \tag{1}$$

This is the pdf of a Cauchy distribution with scale parameter $\gamma/2$; in spectroscopy γ is called the *peak-width at half-height* (or sometimes the *linewidth*).

Each spectral peak corresponds to magnetic nuclei resonating at a particular frequency in the biological mixture. This frequency determines the displacement of the peak on the x-axis, which is known as its *chemical shift* and is measured in parts per million (ppm) of the resonant frequency of a standard peak. It is conventional in NMR spectroscopy to use δ to denote chemical shift and for the δ-axis to increase from right to left.

^{1}H NMR only detects the resonance of hydrogen nuclei and a typical ^{1}H NMR spectrum has a range of about 0ppm-10ppm.

The resonant frequencies of a magnetic nucleus are largely determined by its molecular environment, that is the chemical structure of the molecule in which it is embedded and the configuration of its chemical bonding within the molecule. Consequently, every metabolite has a characteristic molecular ^{1}H NMR *signature*, a convolution of Lorentzian peaks that appear in specific positions in ^{1}H NMR spectra. These are the peaks observed in the ^{1}H NMR spectrum of a pure solution of the metabolite. The peaks of a signature can have quite different chemical shifts (when they are generated by protons with different bonding configurations) and so appear widely separated in a spectrum.

Depending on its molecular environment, a proton may have more than one resonant frequency and when this happens the frequencies are usually very similar. Consequently, the peaks generated appear in a spectrum as a juxtaposition called a *multiplet*. The shape of a multiplet (number of peaks, their separations and relative heights) can be used to identify the corresponding metabolite. Figure 2 shows an ^{1}H NMR spectrum (top panel) and the resonance signatures of the four metabolites contributing the principal resonance signals (lower panels), with characteristic peak locations and multiplet shapes.

[Figure 2 about here.]

The intensity of a nuclear magnetic signal is proportional to the number of magnetically equivalent nuclei generating the resonance in the biological mixture. Consequently, every resonance peak (and therefore also every metabolite resonance signature) scales vertically in a spectrum in proportion to the molecular abundance of the corresponding compound in the mixture.

1.2 Specific challenges of NMR in metabolomics

Biofluids and tissue samples usually contain thousands of metabolites. However, because of poor signal to noise ratios at low concentrations, NMR is relatively insensitive, so that ordinarily a spectrum contains quantitative information on just a few hundred of the most abundant compounds. These compounds can generate hundreds of resonance peaks in a spectrum, many of which overlap.

To quantify a collection of metabolites using NMR, at least one resonance peak generated by each compound must be identified in the spectrum and deconvolved. (To reduce uncertainty, it is desirable to identify as many peaks as possible for each compound.) Estimates of the relative concentrations of the metabolites in the biological sample can be made by comparing the areas under the deconvolved resonance peaks (estimates of absolute concentration need a reference).

The peak identification step (*assignment*) is complicated by fluctuations in peak positions between spectra, caused by uncontrollable differences in experimental conditions and differences in the chemical properties of the biological samples, such as the pH and ionic strength. When this *positional noise* combines with peak overlap, assignment can become very hard indeed.

[Figure 3 about here.]

The left panel of Figure 3 illustrates the problem. Excerpts from two spectra corresponding to biological replicates from the same experiment are overlaid, focusing on a doublet type multiplet with two peaks. The difference in peak position between replicates is obvious to the eye. Here, the magnitude of the positional noise is insufficient to confuse assignment by an expert spectroscopist but it will pose problems for standard automated approaches. However, expert deconvolution is rarely practical because it is labor intensive and relies on someone familiar with metabolite resonance patterns.

Targeted profiling (Weljie et al. 2006) against a standard library of metabolite resonance peaks reduces the importance of expert knowledge but is slow because there is no automated fitting procedure. Spectral binning (Holmes et al. (1994), Spraul et al. (1994)) approaches divide the spectrum into regions (bins), within which the intensity measurements are averaged, in an attempt to isolate distinct resonance signals. Although this mitigates the effect of peaks fluctuating position within bins, fluctuations across bin boundaries will cause anti-correlated increase/decrease of average intensity in adjacent bins, even if there is no associated change in metabolite concentration. Spectral binning balances parsimony and computational efficiency, retaining quantitative information but in a representation with many fewer, easy to compute, variables. However, the reduced variables are often analyzed without explicit quantification of individual metabolites, using pattern recognition methods such as principal components analysis and partial least squares (Lindon, Holmes and Nicholson 2001).

The additional complication of positional noise combined with peak overlap is illustrated in the right panel of Figure 3. The well defined resonance peaks overlap with broad signals attributable to a combination of closely overlapping low metabolite signals and/or macromolecular signatures. This introduces the problem of estimating the proportion of the signal associated with the sharp resonances and the proportion due to the broad component. The problem becomes even more complex when the target peaks also overlap with other sharp metabolite signals which additionally fluctuate between different spectra.

Currently, there is no statistical methodology that can simultaneously address the problems of identification, deconvolution and quantification when there is positional noise and peak overlap. We believe a method based on explicit quantification from deconvolution of metabolite signatures should have significant advantages: spectral convolution models are parsimonious because they correspond to the physical process

generating the data; the variables inferred are interpretable because they represent concentrations of identified metabolites; these concentrations are of direct scientific interest because they depend on the underlying biology.

1.3 Contributions of this paper

To tackle the problem of quantifying metabolites in complex biofluids such as cell supernatants or urine, we present a Bayesian model for ^{1}H NMR spectra and a Markov chain Monte Carlo (MCMC) algorithm to automate peak assignment and spectral deconvolution. Bayesian models for NMR data have been described before, notably in Bretthorst (1990a), in Bretthorst (1990b), in many subsequent papers by the same author, in Dou and Hodgson (1996) and in Rubtsov and Griffin (2007). Our modeling exploits extensive prior information on the resonance signatures of the metabolites, including the expected horizontal displacements and relative vertical scalings of the peaks. This novel approach allows us to deconvolve peaks and assign them to specific metabolites in a unified analysis, which eliminates the need for a manual assignment step. The prior information comes from the physical theory of NMR and from experimental information. Experimental resonance data on human metabolites are extensive and are publicly available, for example from the online database of the Human Metabolome Project, the HMDB (Wishart et al. 2009).

Almost all biofluid and tissue NMR spectra contain peaks for which there is no prior information in the presently incomplete public metabolite databases. Despite this, the component of a spectrum that cannot be assigned to known compounds, may contain metabolomic information that is scientifically useful, e.g. for classification of spectra. We therefore propose a two component joint model for a spectrum. We model the metabolites whose peaks we wish to assign explicitly parametrically, using information from the online databases, while we model the unassigned spectrum semi-parametrically, using wavelets. We choose wavelets because they model

signal continuously but locally. They can account for the local correlation of a spectrum caused by the continuity of the underlying physical processes without imposing unrealistic global modeling constraints.

The wavelet component of our two component likelihood is extremely flexible so that, without restriction, it tends to absorb signal that should be modeled by the parametric component, thus inducing a lack of identifiability. We address this by penalizing the wavelet coefficients using heavy tailed scale-mixture priors. These priors shrink wavelet coefficients wherever the spectral signal can be explained by the parametric component of the model. We also impose a truncation condition on the wavelets, which reflects prior knowledge that frequency-domain NMR spectra lie almost completely in the upper-half of the (x, y) plane.

To overcome the strong posterior correlation between parameters corresponding to the two model components we introduce purposely designed Metropolis-Hastings block proposals which update the parameters of the two components jointly.

2. METHODOLOGY

2.1 Modeling NMR Spectra

Previous authors (Bretthorst (1990a), Dou and Hodgson (1996), Rubtsov and Griffin (2007)) developed Bayesian models for NMR data in the time domain, in which resonance signals appear as exponentially decaying sinusoids. However, we prefer to model conventionally preprocessed (by apodization, phase and baseline correction) data in the more interpretable frequency domain, in which resonance signals appear as peaks (e.g. Figure 1). Our model exploits the positivity of the frequency-spectrum, a condition which cannot be expressed parsimoniously in the time domain. Under an iid Gaussian model for errors, the two representations contain the same information since they are related by an orthogonal transformation (the discrete Fourier transform).

A frequency domain NMR dataset is a pair $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}$ is a length n vector of points on the chemical shift axis, usually regularly spaced and $\mathbf{y}$ is a vector of corresponding resonance intensity measurements. The intensity measurements are noisy, so that, although they measure inherently positive quantities, some components of $\mathbf{y}$ are likely to fall below the δ-axis. $\mathbf{y}$ is usually standardized in some way, for example so that $\sum_i^n y_i = 1$.

We model $\mathbf{y}|\mathbf{x}$ assuming the $y_i|\mathbf{x}$ are independent normal random variables and,

$$\mathbb{E}\left(y_i|\mathbf{x}\right) = \phi(x_i) + \xi(x_i) \tag{2}$$

where the ϕ component of the model represents signal from metabolites with peaks we wish to assign explicitly and which have been previously characterized and cataloged in online databases. (The metabolites chosen will vary from analysis to analysis according to the prior belief about the content of the biological mixture and the scientific question.) The ξ component of the model represents signal generated by peaks we do not wish to assign (this may include signal from uncatalogued resonances of molecules which are partially characterized with the characterized resonances modeled in the ϕ component). We construct ϕ parametrically (as a continuous function of continuous chemical shift δ) using the physical theory of NMR and we model ξ semi-parametrically using wavelets.

2.2 Modeling ϕ, the cataloged metabolite signal

According to physics, the resonance signatures of distinct compounds are independent, accumulate with an intensity proportional to molecular abundance and aggregate in a spectrum by convolution. Consequently, we can write ϕ as a linear combination, where each term in the sum corresponds to the signature of one of M different metabolites,

$$\phi(\delta) = \sum_{m=1}^{M} t_m(\delta)\beta_m. \tag{3}$$

For each m, t_m is a continuous template function which specifies the NMR signature of metabolite m. The corresponding coefficient β_m is proportional to the molecular abundance of m (i.e. the concentration of m) in the biological sample. The physics implies that the template functions have a particular parametric form; we will describe a parametric prior for the t_m, in detail, below.

2.3 Modeling ξ, the uncatalogued metabolite signal

We model $(\xi(x_1), ...\xi(x_n))^T$ as a linear combination of wavelet basis functions and use $\boldsymbol{\theta}$ to denote the vector of wavelet coefficients. We chose to use the symlet-6 wavelet basis because these wavelets have a similar shape to Lorentzian peaks. Symlets have been used previously to select features from NMR spectra (Kim et al. 2008) and sensitivity analysis comparing other potential wavelet bases showed little difference in spectral reconstructions.

2.4 The Likelihood

We now bring together the models for ϕ and ξ to make a formal specification of the probability model for the data. It is easier to do this in the wavelet domain, because the dimension of the wavelet space p often needs to be greater than the dimension of the data space n, to deal with distortion at the spectral borders (see Strang and Nguyen (1996) and Section 1 of the supplementary material.) Let $\mathcal{W}$ be the wavelet transform corresponding to the symlet-6 wavelet basis. The likelihood, is defined by

$$\mathcal{W}\mathbf{y} = \mathcal{W}\mathbf{T}\boldsymbol{\beta} + \mathbf{I}_p\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \mathbf{I}_p/\lambda), \tag{4}$$

where $\mathbf{T}$ is the $n \times M$ matrix with $t_m(x_i)$ as its (i,m)th entry, $\mathbf{I}_p$ is the $p \times p$ identity matrix and where λ is a scalar precision parameter.

Equation (4) is a linear regression of $\mathcal{W}\mathbf{y}$ on the columns of $[\mathcal{W}\mathbf{T}\ \mathbf{I}_p]$, the matrix generated by adjoining $\mathcal{W}\mathbf{T}$ and $\mathbf{I}_p$ columnwise. Since this matrix has more columns than rows, the regression coefficients cannot all be identifiable in the likelihood. We address this in the next section, by specifying a prior which helps to distinguish the parametric and semi-parametric components of the model.

2.5 Prior specification

Our aim is to obtain a joint Bayesian posterior distribution over the parameters $\{t_m : m = 1, ..., M\}$, $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and λ; we now specify priors for these parameters.

Prior for t_m, the metabolite signature templates : The physics restricts the model space for each template to a parametric mixture of horizontally translated and vertically Lorentzian peaks (Hore 1995). Our focus is on spectra generated by biofluids such as cell supernatants or urine, for which peak-widths vary between, but negligibly within spectra; we therefore make the assumption that peaks within a spectrum depend on a single common peak-width parameter γ. (This assumption could easily be relaxed, to allow a local deviation at each Lorentzian peak, by including random effects.) We specify a translated and scaled beta prior distribution for γ, with both shape parameters equal to 2 and with support $[0.7\text{Hz}/F, 1.3\text{Hz}/F]$, where F is the operating frequency of the spectrometer in MHz. This prior is quite uninformative in the region around $1\text{Hz}/F$, which is typical of the peak-widths generated by modern spectrometers (Hore 1995).

As described above, many metabolite signatures contain clusters of peaks called multiplets. Isolated peaks are often classed as *singlet* multiplets and, if we adopt this convention, we can express each signature template t_m completely as a linear

combination of a set of multiplet curves g_{mu},

$$t_m(\delta) = \sum_u z_{mu} g_{mu}(\delta - \delta^\star_{mu}), \tag{5}$$

where u is an index running over all the multiplets belonging to metabolite m. We assume $\int_0^\infty g_{mu}(\delta)d\delta = \int_{-\infty}^0 g_{mu}(\delta)d\delta$ for all m, u, so that the parameter $\delta^\star_{mu}$ specifies the position on the chemical shift axis of the center of mass of the uth multiplet of the mth metabolite (see large multiplet in Figure 4). We call $\delta^\star_{mu}$ the chemical shift parameter of the multiplet. Each of the coefficients z_{mu} is a positive integer equal to the number of protons in a molecule of m that contribute resonance signal to the multiplet u. $\int_{-\infty}^\infty g_{mu}(\delta)d\delta$ is constant over m and u so the area under each t_m is proportional to $\sum_u z_{mu}$, the number of protons resonating in a molecule of m.

With a few exceptions, most multiplets can be classified into one of a number of common types (Figure 4) which determine the configuration of the peaks (a doublet, a triplet, a doublet of doublets, etc). This classification together with a small number of continuous quantities called J-coupling constants, which determine the (horizontal) distances between the peaks, completely parameterize a multiplet curve.

[Figure 4 about here.]

To be precise, a multiplet curve g_{mu} is the weighted average of V_{mu} translated Lorentzian curves (see eqn. (1)),

$$g_{mu}(\delta) = \sum_{v=1}^{V_{mu}} w_{muv} l_\gamma(\delta - c_{muv}), \tag{6}$$

where the weights w_{muv} (which sum to one over v) determine the relative heights of the peaks of the multiplet and the translations c_{muv} determine the horizontal offsets of the peaks from the center of mass of the multiplet (see Figure 4). Multiplets

are (usually) symmetric so that $\{-c_{muv} : v = 1...V_{mu}\} = \{c_{muv} : v = 1...V_{mu}\}$ and $w_{muv'} = w_{muv}$ when $c_{muv'}{=}{-}c_{muv}$.

We now have a two-level parameterization of signature templates, defined by (5) and (6), as a linear combination of Lorentzian peaks nested in multiplets. This allows us to represent a difference in the uncertainty of peak positions within and between multiplets. The parameters c_{muv} and w_{muv}, which determine the multiplet shapes, vary very little across NMR spectra. We assume they are constant and compute them by applying some simple rules (see Hore (1995), chap. 3 for the details), from empirical estimates of the J-coupling constants which are published in online databases. In contrast, as discussed in section 1.2, the multiplet chemical shift parameters $\delta^{\star}_{mu}$ do fluctuate slightly between spectra according to experimental conditions. We use an estimate $\hat{\delta^{\star}}_{mu}$ of each $\delta^{\star}_{mu}$, taken from online databases, to construct an informative prior which accounts for this uncertainty. The positional noise is local and smaller fluctuations are more probable than larger ones, so we assign each $\delta^{\star}_{mu}$ a truncated normal prior distribution with mean parameter $\hat{\delta^{\star}}_{mu}$, variance parameter 10^{-4}ppm^2 and truncation region $[\hat{\delta^{\star}}_{mu} - 0.03\text{ppm}, \hat{\delta^{\star}}_{mu} + 0.03\text{ppm}]$.

Prior for the abundance parameters $\boldsymbol{\beta}$: Having defined a parametric prior for the metabolite signature templates, we now consider the prior for the vector $\boldsymbol{\beta}$, each component of which represents the resonance intensity of a signature and is proportional to the abundance of that metabolite in the biological sample. The intensities are positive so the prior for each component of $\boldsymbol{\beta}$ should confine its support to $\mathbb{R}^{+}$. For conditional conjugacy, we assign a normal prior to each β_m, truncated below at zero, $\beta_m \sim TN(e_m, 1/s_m^2, 0, \infty)$. This is flexible enough to encode prior information for a wide range of research problems. For the simulations and examples presented in this paper we assume low prior information and choose $e_m = 0$ and $s_m^2 = 10^{-3}$ for all m.

Prior for the wavelet parameters: In section 2.4 we observed that the parametric (ϕ) and semi-parametric (ξ) components of the model are not identifiable in the likelihood. In order to resolve the model components we penalize the semi-parametric component by assigning each wavelet coefficient a prior distribution with a concentration of probability mass near zero. In addition, to reflect prior-knowledge that NMR spectra are mostly restricted to the half-plane above the chemical shift axis, we use a prior which penalizes models in which $\mathcal{W}^{-1}\boldsymbol{\theta}$ has components below a small negative threshold.

Figure 5 demonstrates the effect of penalizing the semi-parametric component of the model when it lies below the chemical shift axis. First, without penalizing ξ in the lower half plane, if the components of $\boldsymbol{\theta}$ are given untruncated Student's-t priors, the posterior for the vector of quantification parameters $\boldsymbol{\beta}$ focuses asymptomatically on a region close to the ordinary least squares estimate of the parameters, $(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y}$, while the wavelet component absorbs most of the residual spectrum. When a metabolite has a multiplet embedded in a region of unassigned spectral resonance the least squares estimate overestimates the corresponding quantification parameter. The signature templates absorb spectral signal even when they do not match the shape of the spectral data and this leads to strong posterior support for negative values for some components of $\mathcal{W}^{-1}\boldsymbol{\theta}$. However, a prior penalizing ξ in the lower half plane, can correct the concentration estimate providing at least one of the multiplets of the metabolite is deconvolved cleanly.

[Figure 5 about here.]

One way to implement a prior for $\boldsymbol{\theta}$ with the properties described above, is to use a scale-mixture of Gaussians prior for each coefficient and then introduce a joint smoothed truncation of the wavelet coefficients. For instance we might specify a joint

prior for $(\boldsymbol{\theta}, \lambda)$ as follows

$$P(\boldsymbol{\theta}|\lambda, \boldsymbol{\psi}, \boldsymbol{\tau}) = \frac{\lambda^{p/2} \prod_{jk} \psi_{jk}^{1/2}}{C_{\lambda\boldsymbol{\psi}\boldsymbol{\tau}}} \exp\left(-\frac{1}{2}\sum_{jk} \lambda\psi_{jk}\theta_{jk}^2\right) \mathbb{1}_{\{\mathcal{W}^{-1}\boldsymbol{\theta} \geq \boldsymbol{\tau}\}}, \tag{7}$$

$$\lambda \sim \text{Gamma}(a, b/2), \tag{8}$$

$$\psi_{jk} \sim \text{Gamma}(c_j, d_j/2), \tag{9}$$

$$\tau_i \sim h - \text{Exp}(r), \tag{10}$$

where $C_{\lambda\boldsymbol{\psi}\boldsymbol{\tau}}$ is a normalizing constant. The index jk here corresponds to the kth wavelet in the jth wavelet-scaling level and the index i corresponds to the ith spectral data point. $\boldsymbol{\psi}$ is a vector of hyperparameters, which allow the prior precision associated with each wavelet to deviate from the global precision λ. The gamma hyperprior on each component of $\boldsymbol{\psi}$ induces a shrinkage distribution as the marginal prior for $\boldsymbol{\theta}$, which encourages posterior sparsity in the wavelet coefficients.

The parameter $\boldsymbol{\tau}$ is a vector of n truncation limits, which bounds $\mathcal{W}^{-1}\boldsymbol{\theta}$ below. The exponential hyperpriors on the components of $\boldsymbol{\tau}$ smooth these truncations and penalize $\boldsymbol{\theta}$ more heavily as more of the semi-parametric component of the model lies below the line $y = h$, where h is a small negative number, chosen close to zero on the spectral intensity scale.

The joint distribution specified by (7)-(10) is a reasonable representation of prior belief about $\boldsymbol{\theta}$ and λ; it places a constraint on the *conditional* distribution of $\boldsymbol{\theta}$ given λ and $\boldsymbol{\tau}$. However, because the normalizing constant $C_{\lambda\boldsymbol{\psi}\boldsymbol{\tau}}$ of (7) has no closed form, it is hard to devise a computationally efficient scheme to sample from the resulting 'doubly intractable' posterior (Murray et al. 2006). We therefore prefer to define a

joint prior for $(\boldsymbol{\theta}, \lambda, \boldsymbol{\psi}, \boldsymbol{\tau})$, with pdf proportional to

$$\lambda^{p/2+a-1}\left[\prod_{jk}\psi_{jk}^{c_j-1/2}\right]\exp\left(-\frac{1}{2}b\lambda-\frac{1}{2}\sum_{jk}\psi_{jk}\left(d_j+\lambda\theta_{jk}^2\right)+r\mathbf{1}_n^T\boldsymbol{\tau}\right)\mathbb{1}_{\{\mathcal{W}^{-1}\boldsymbol{\theta}\geq\boldsymbol{\tau}\}}\mathbb{1}_{\{h\mathbf{1}_n\geq\boldsymbol{\tau}\}} \tag{11}$$

where $\mathbf{1}_n$ is the column vector with n components, all equal to 1. This prior places the constraint on the *joint* distribution of the parameters, rather than the conditional distribution; we contend that this is an equally valid specification. It is easy to sample from the full conditionals of $\boldsymbol{\theta}$, λ, $\boldsymbol{\psi}$ and $\boldsymbol{\tau}$ if we use this prior and we show in Section 2 of the supplementary material that it behaves similarly to the prior defined by (7)-(10). For simulations and examples described here we take $a = 10^{-9}$, $b = 10^{-6}$, $c_j = 2$, $d_j = 400 \times 1.05^{-j}$, $h = -0.01$ and $r = 4 \times 10^2$.

3. MARKOV CHAIN MONTE CARLO ALGORITHM

We implemented a Markov chain Monte Carlo algorithm, to sample from the joint posterior distribution of the model parameters. During the burn in stage of the MCMC we temper the likelihood and penalize the wavelet component of the model stringently to move the chain into a region of good posterior support.

The updates are described in detail in Section 3 of the supplementary material. Briefly, we use Gibbs sampling to update λ and to update the components of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, $\boldsymbol{\psi}$ and $\boldsymbol{\tau}$. We use Metropolis-Hastings updates with Gaussian proposals, centered on the current parameter value and truncated at the boundaries of the prior distributions, to update the multiplet chemical shift parameters $\delta_{mu}^{\star}$ and the peak-width parameter γ. We adapt these proposals using the Adaptive Metropolis-within-Gibbs algorithm of Roberts and Rosenthal (2009).

The updates summarized above allow the Markov chain to explore the posterior locally, with $\boldsymbol{\beta}$ and the $\delta_{mu}^{\star}$ mixing slowly. However, the chain cannot make large moves in the parameter space with these updates because the likelihood (4) constrains

$\mathcal{W}\mathbf{y} - \mathcal{W}\mathbf{T}\boldsymbol{\beta} - \boldsymbol{\theta}$, inducing strong posterior correlation between $\boldsymbol{\theta}$ and the parameters controlling $\mathcal{W}\mathbf{T}\boldsymbol{\beta}$ (i.e. $\boldsymbol{\beta}$, γ and $\delta^{\star}_{mu}$). To allow the chain to make global moves we introduce block updates for each component of $\boldsymbol{\beta}$ jointly with $\boldsymbol{\theta}$ and for each $\delta^{\star}_{mu}$ jointly with $\boldsymbol{\theta}$.

We will just describe the block update of each β_m with $\boldsymbol{\theta}$ here. (The block updates for each $\delta^{\star}_{mu}$ with $\boldsymbol{\theta}$ follow a similar format and the technical details are in Section 3 of the supplementary material.) First, we propose β'_m, a new value for β_m, from a Cauchy distribution, truncated below at zero. We center the Cauchy distribution on the point that maximizes the full conditional of β_m subject to $\boldsymbol{\theta} = \mathbf{0}$ and to the truncation condition $\mathbf{y} - \mathbf{T}\boldsymbol{\beta} > \boldsymbol{\tau}$. Next, conditional on β'_m, we propose a new value $\boldsymbol{\theta}'$ for $\boldsymbol{\theta}$ and perform a Metropolis-Hastings accept-reject for $(\boldsymbol{\beta}', \boldsymbol{\theta}')$. The conditional proposal for $\boldsymbol{\theta}'$ is a multivariate truncated normal distribution with mean parameter $\mathcal{W}\mathbf{y} - \mathcal{W}\mathbf{T}\boldsymbol{\beta}'$ (where $\boldsymbol{\beta}'$ is $\boldsymbol{\beta}$ with the mth component set to β'_m), precision parameter $\lambda\mathbf{I}_p$ and truncation $\mathcal{W}^{-1}\boldsymbol{\theta}' \geq \boldsymbol{\tau}$. We can simulate from this distribution by making the change of basis, $\boldsymbol{\eta}' = \mathcal{W}^{-1}\boldsymbol{\theta}'$; the components of $\boldsymbol{\eta}'$ are then independent univariate truncated normal distributions. The choice of proposal for $\boldsymbol{\theta}'$ is motivated by the full conditional of $\boldsymbol{\theta}$ which, (subject to the update of β_m) has a similar distribution but with a reduced precision parameter: the precision parameter of the full conditional is also diagonal but has $\lambda(1 + \psi_{jk})$ in the position corresponding to the jkth wavelet. Unfortunately there is no easy way to sample from this distribution because the truncation condition $\mathcal{W}^{-1}\boldsymbol{\theta} \geq \boldsymbol{\tau}$ induces a complex dependence structure between the components of $\boldsymbol{\theta}$.

4. PERFORMANCE

We used MetAssimulo (Muncey et al. 2010), a MATLAB package for NMR spectral simulation, to generate datasets for assessing our model and sampling algorithm. MetAssimulo simulates biofluid spectra by convolving a library of standard spectra

generated from the NMR of pure metabolite samples. The software also simulates multiplet positional noise.

4.1 Multiplet Localization

We simulated biofluid spectra with metabolite concentrations typical of normal human urine, except that we bounded the simulated concentration of malic acid to ensure the metabolite was present at a detectable level in each spectrum. Malic acid generates three 'doublet of doublet' type multiplets in an NMR spectrum at 2.35ppm, 2.66ppm and 4.29ppm. For each of the 3 multiplets, we simulated 21 urine spectra, varying the chemical shift of the multiplet on a grid of points within 0.03ppm (18Hz) of the chemical shift estimate published in the HMDB.

We applied the MCMC algorithm to each simulated urine spectrum, with the parametric component of the model composed of the single template corresponding to malic acid. The results are summarized in Figure 6. We obtained excellent localization of the 3 multiplets across the range of simulated chemical shifts, except in three simulations. These failures all occurred in simulations for which the concentration of malic acid was low (in the bottom 15% of simulated concentrations) and where the signal from the target multiplet was close to or heavily convolved with a strong resonance from another compound, which influenced the alignment.

[Figure 6 about here.]

4.2 Quantification

We compared our Bayesian method for quantification to the following algorithm for numerical integration, which is slightly more sophisticated than spectral binning because it includes a peak identification step. To integrate multiplet u of metabolite m, we estimate the chemical shift of the multiplet by identifying spectral peaks in the region $[\hat{\delta^\star}_{mu} - 0.03\text{ppm}, \hat{\delta^\star}_{mu} + 0.03\text{ppm}]$ (i.e. the support of the prior for $\delta^\star_{mu}$) and

matching them to the multiplet. (The multiplets generated by these simulations all have a central maximum so we simply match the spectral data point with the greatest intensity value to the center of the multiplet.) We then identify the region $[L, R]$ of the chemical shift axis that corresponds to the central 95% mass of the multiplet's parametric template (6) (e.g. center ± 0.023ppm for a singlet Lorentzian). Finally, we estimate the total area under the multiplet by $\left(\sum_{L<x_i<R} y_i\right) / (0.95N(L-R))$, where N is the number of x_i in $[L, R]$.

Using MetAssimulo, we generated NMR spectra typical of ordinary human urine, varying the simulated concentrations of two compounds, 2-oxoglutaric acid and dimethylglycine, across the simulated spectra. Each of these metabolites generate two multiplets in NMR spectra; 2-oxoglutaric acid generates two triplets (at 2.42ppm and 2.99ppm); dimethylglycine generates two singlets (at 2.91ppm and 3.71ppm).

We generated 64 spectra, each corresponding to a point in an 8×8 grid of concentration pairs. The simulated concentrations were chosen so that the resonances generated by the two metabolites were distinguishable from the background noise and comparable in magnitude to other distinguishable resonances. Figure 7 shows a typical simulated spectrum.

[Figure 7 about here.]

Because each spectrum in the MetAssimulo library was generated from a different ^{1}H NMR experiment, the width of the peaks in a simulated spectrum can vary slightly between resonances generated by different compounds. This variability can affect inferences of concentration made with our model because of correlation between concentration and peak-width parameters. To deal with this artifact of the simulator, we estimated the peak-width parameter separately for the two metabolites using their library spectra and fixed the peak-widths of their resonances in our model.

We ran the MCMC algorithm on each simulated dataset for 5000 iterations following a 5000 iteration burn in. Figure 8 summarizes the posterior distributions of the concentration parameters across the simulations. The posterior mean tends to overestimate the concentrations of both compounds slightly, presumably because the template component of the model absorbs some signal from convolved background resonances. The bias is a little more pronounced at low concentrations when it is harder to distinguish the signal precisely from the background. The Monte Carlo estimate of the 95% credible region covers the simulated concentration pair for 58 of the 64 simulations.

Figure 9 is a comparison of the bias in the concentration estimates (estimated-simulated) for the posterior mean of the Bayesian model and for the numerical integration method. Despite the slight error the Bayes estimates provide a striking improvement on numerical integration. On average, both methods overestimate concentration, but the mean absolute bias in the Bayesian estimates are, respectively for 2-oxoglutaric acid and dimethylglycine, 11.5 and 11.2 fold lower than the mean absolute bias for the numerical integration. Both methods are better at estimating the concentration of dimethylglycine than the concentration of 2-oxoglutaric acid. This is because the multiplets of dimethylglycine generally do not overlap other signal, while the triplet of 2-oxoglutaric acid at 2.42ppm is often convolved with a multiplet of succinate which appears at 2.41ppm and the triplet of 2-oxoglutaric acid at 2.99ppm is often convolved with a singlet of creatinine, which appears at 3.03ppm (e.g. see Figure 7).

[Figure 8 about here.]

[Figure 9 about here.]

5. YEAST DATASET

To illustrate the performance of our method on a real dataset, we took three spectra from the experiment investigating the metabolic response of the yeast *Pichia pastoris* to recombinant protein expression (Tredwell et al. 2011). The spectra were generated using biological replicates prepared under the same conditions. Because of this, the metabolic profiles of the samples are extremely similar and the spectra contain essentially the same metabolite concentration information. Nevertheless, the spectra are slightly different, because for example, of experiment level positional noise in the chemical shifts of resonance peaks. By modeling the three spectra jointly we can quantify metabolites using information from all three replicates, while accounting for these experiment level differences.

We used the model described in Section 2 as a basis for a joint model of multiple spectra. In the new model, the vector of metabolite quantification parameters $\boldsymbol{\beta}$ is held in common across the spectra. All the remaining parameters are copied from the original model, with a replicate set assigned to each spectrum. The MCMC algorithm for the multiple-spectra model is very similar to the procedure described for the original model. The Metropolis-Hastings updates involving components of $\boldsymbol{\beta}$ need to be adjusted, to reflect the dependence on multiple spectra, but are similar to those for the simpler model (see Section 3 of the supplementary material). The updates for the remaining parameters continue to be valid within each spectrum because, conditional on $\boldsymbol{\beta}$, the joint posterior factorizes into separate probability models, each corresponding to a different spectrum.

We picked seven metabolites, alanine, arabitol, malic acid, methionine sulfoxide, threonine, trehalose and succinic acid thought to be present in the yeast samples at non-trivial concentrations. Not all the resonance patterns generated by these metabolites take the form of the symmetric multiplets described in Section 2. Mul-

tiplet shapes are sometimes distorted by strong interaction effects, which we cannot model because they are not cataloged in a public database. Consequently, we were unable to build complete parametric signature templates for arabitol and trehaolose but were able to build approximate templates by omitting the interacting multiplets. Omission of the multiplets precludes a complete deconvolution of the spectral signal generated by each of the seven compounds since omitted resonance signals will be absorbed into the wavelet component of the model. However, our main aim is to obtain accurate concentration estimates and this is still achievable, providing at least one multiplet from each metabolite deconvolves correctly.

We ran the MCMC procedure for the joint model over 30 000 iterations following a 10 000 iteration burn in. In Figure 10 we show the posterior deconvolution of one of the spectra for the resonances of four of the seven metabolites. (The deconvolution of each spectrum into the resonances of all seven metabolites is shown in Section 4 of the supplementary material.)

[Figure 10 about here.]

We compared our deconvolution to the manual deconvolution of an experienced NMR spectroscopist. Even manual assignment by an expert is subject to error and uncertainty. However, except for some multiplets of malic acid, whenever the spectroscopist could identify and manually deconvolve a spectral resonance, our deconvolution agreed with his. Malic acid did not deconvolve as well as the other metabolites in any of the three spectra. Figure 11 shows the posterior mean deconvolution from the Bayes model with the chemical shift parameters set to MAP estimates and also shows the deconvolution with chemical shift parameters set to the estimates of a spectroscopist. The multiplet that appears at 4.30ppm in the expert deconvolution, is offset in the Bayesian deconvolution because the model prefers to absorb the signal from the peak at 4.27ppm into the parametric component of the likelihood, rather than

model it with wavelets which are heavily penalized. The multiplet which appears at 2.36ppm behaves similarly. This is a weakness of the modeling approach, which would be resolved were the parametric template for the metabolite generating the peak at 4.27ppm included in the model. Despite the errors in the Bayes deconvolution the concentration estimate for malic acid is very close to the estimate derived from expert assignment.

[Figure 11 about here.]

6. DISCUSSION

Presently, automatic methods for analyzing biofluid NMR data rely on non-parametric pattern recognition techniques or are based on approximate numerical integration algorithms, such as spectral binning. These methods ignore a large amount prior information about the physical process generating the spectral data.

Prior information about a data generating process can easily be incorporated into a Bayesian analysis through specification of a likelihood and specification of a prior distribution for the parameters of the likelihood. We have shown that a Bayesian model for biofluid spectra, which exploits an informative parametric prior for the patterns of resonance generated by selected metabolites, can be used to deconvolve those resonances from a spectrum and to obtain explicit concentration estimates for the metabolites.

Simulations show that our MCMC algorithm usually identifies spectral resonance peaks precisely, with occasional mistakes at low concentrations. Specifically, when the target resonances for a multiplet of a template appear in the spectrum close to other stronger signals, the model may encourage the template to align incorrectly with the stronger signals, even if they have the wrong shape. This is because the wavelet coefficients are heavily penalized in the prior but the parametric templates

are not.

It is worth noting that resonance mis-assignment is a problem for all methods, (including manual assignment by an expert; it is unavoidable for binning methods when peaks overlap) and in our approach mistakes can be resolved by adding the signature templates corresponding to the compounds generating the confounding signals to the parametric component of the model (providing they are available). Furthermore, even when the model posterior concentrates around an incorrect deconvolution the strong prior penalization on negative spectral signal means that posterior estimates of concentration can still be accurate, providing at least one multiplet for each metabolite deconvolves correctly. Concentration estimation, the main motivation for the modeling, is therefore quite robust to mis-assignment of spectral resonances.

Our approach yields improved concentration estimates. A comparison with a method for quantifying metabolites based on numerical integration shows the posterior mean estimates of the Bayesian model to be 11 fold more accurate, even when the conditions are favorable for numerical integration, i.e. when the targeted peaks are distinct and do not overlap other resonances.

An accurate, automatic method for estimating metabolite concentrations from ^{1}H-NMR spectra will assist many research projects in metabolomics. The field relies heavily on NMR for metabolite quantification and currently, even projects analyzing a few tens of spectra use numerical integration for estimation. Bayesian modeling should become increasingly useful as prior information on metabolite resonance patterns becomes more accurate and extensive. With more detailed information our template model could be extended to include, for example, the effects of interactions between multiplets. We plan to develop our model further and to release an efficient implementation of our methodology capable of simultaneously deconvolving the majority of metabolites assignable in the NMR spectra of complex biological mixtures.

REFERENCES

Bretthorst, G. L. (1990a), "Bayesian Analysis. I. Parameter Estimation Using Quadrature NMR Models," Journal of Magnetic Resonance, 88, 533–551.

Bretthorst, G. L. (1990b), "Bayesian Analysis. II. Signal Detecion and Model Selection," Journal of Magnetic Resonance, 88, 552–570.

Brindle, J., Antti, H., Holmes, E., Tranter, G., Nicholson, J., Bethell, H., Clarke, S., Schofield, P., McKilligin, E., Mosedale, D., and Grainger, D. J. (2002), "Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using ^{1}H-NMR-based metabonomics," Nature Medicine, 8(12), 1439–1445.

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(1-3), 115–20.

Dou, L., and Hodgson, R. (1996), "Bayesian inference and Gibbs sampling in spectral analysis and parameter estimation: II," Inverse Problems, 12, 121–137.

Griffiths, J. R., McSheehy, P. M. J., Robinson, S. P., Troy, H., Chung, Y.-L., Leek, R. D., 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.

Holmes, E., Foxall, P., Nicholson, J., Neild, G., Brown, S., Beddell, C., Sweatman, B., Rahr, E., Lindon, J., Spraul, M. et al. (1994), "Automatic data reduction and pattern recognition methods for analysis of 1H nuclear magnetic reso-

nance spectra of human urine from normal and pathological states," Analytical Biochemistry, 220(2), 284–296.

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(7193), 396–400.

Hore, P. (1995), Nuclear Magnetic Resonance, Vol. 32 of Oxford Chemistry Primers Oxford University Press.

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. (2010), "A genome-wide perspective of genetic variation in human metabolism.," Nature Genetics, 42(2), 137–41.

Kim, S. B., Wang, Z., Oraintara, S., Temiyasathit, C., and Wongsawat, Y. (2008), "Feature selection and classification of high-resolution NMR spectra in the complex wavelet transform domain," Chemometrics and Intelligent Laboratory Systems, 90(2), 161–168.

Lindon, J. C., Holmes, E., and Nicholson, J. K. (2001), "Pattern recognition methods and applications in biomedical magnetic resonance," Progress in Nuclear Magnetic Resonance Spectroscopy, 39(1), 1–40.

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), 137–46.

Muncey, H., Jones, R., De Iorio, M., and Ebbels, T. (2010), "MetAssimulo: simulation of realistic NMR metabolic profiles," BMC Bioinformatics, 11(1), 496.

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006), MCMC for Doubly-intractable Distributions,, in UAI, AUAI Press.

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(1), 45–50.

Roberts, G., and Rosenthal, J. (2009), "Examples of adaptive MCMC," Journal of Computational and Graphical Statistics, 18(2), 349–367.

Rubtsov, D. V., and Griffin, J. L. (2007), "Time-domain Bayesian detection and estimation of noisy damped sinusoidal signals applied to NMR spectroscopy.," Journal of Magnetic Resonance, 188(2), 367–79.

Spraul, M., Neidig, P., Klauck, U., Kessler, P., Holmes, E., Nicholson, J., Sweatman, B., Salman, S., Farrant, R., Rahr, E. et al. (1994), "Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples," Journal of Pharmaceutical and Biomedical Analysis, 12(10), 1215–1225.

Strang, G., and Nguyen, T. (1996), Wavelets and Filter Banks Wellesley Cambridge Press.

Tredwell, G. D., Edwards-Jones, B., Leak, D. J., and Bundy, J. G. (2011), "The Development of Metabolomic Sampling Procedures for Pichia pastoris, and Baseline Metabolome Data.," PLoS One, 6(1), e16286.

Weljie, A. M., Newton, J., Mercier, P., Carlson, E., and Slupsky, C. M. (2006), "Targeted profiling: quantitative analysis of ^{1}H NMR metabolomics data.," Analytical Chemistry, 78(13), 4430–42.

Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., Hau, D. D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J. A., Lim, E., Sobsey, C. A., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H. J., and Forsythe, I. (2009), "HMDB: a knowledgebase for the human metabolome.," Nucleic Acids Research, 37(Database issue), D603–10.

List of Figures

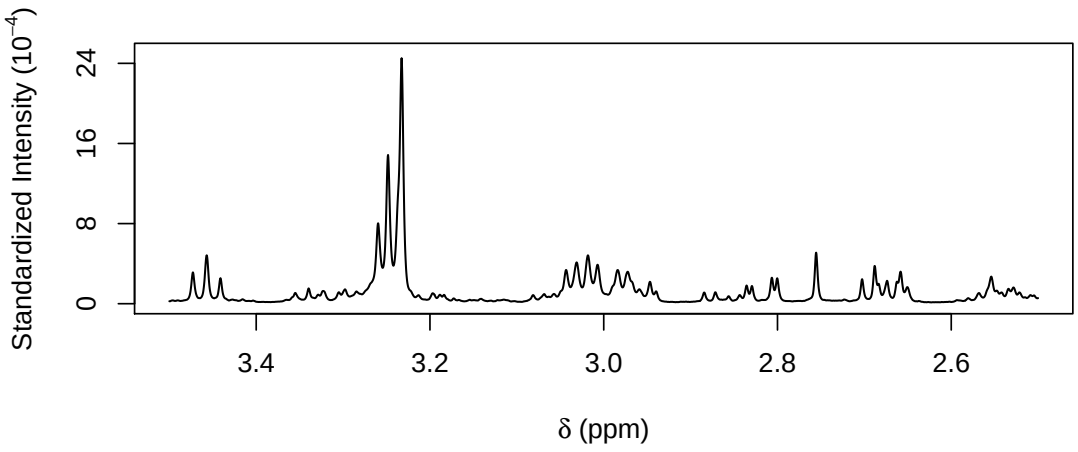


Figure 1: A section from an NMR spectrum from an experiment investigating protein expression in yeast. The x-axis measures chemical shift in parts per million (ppm); the y-axis measures relative resonance intensity.

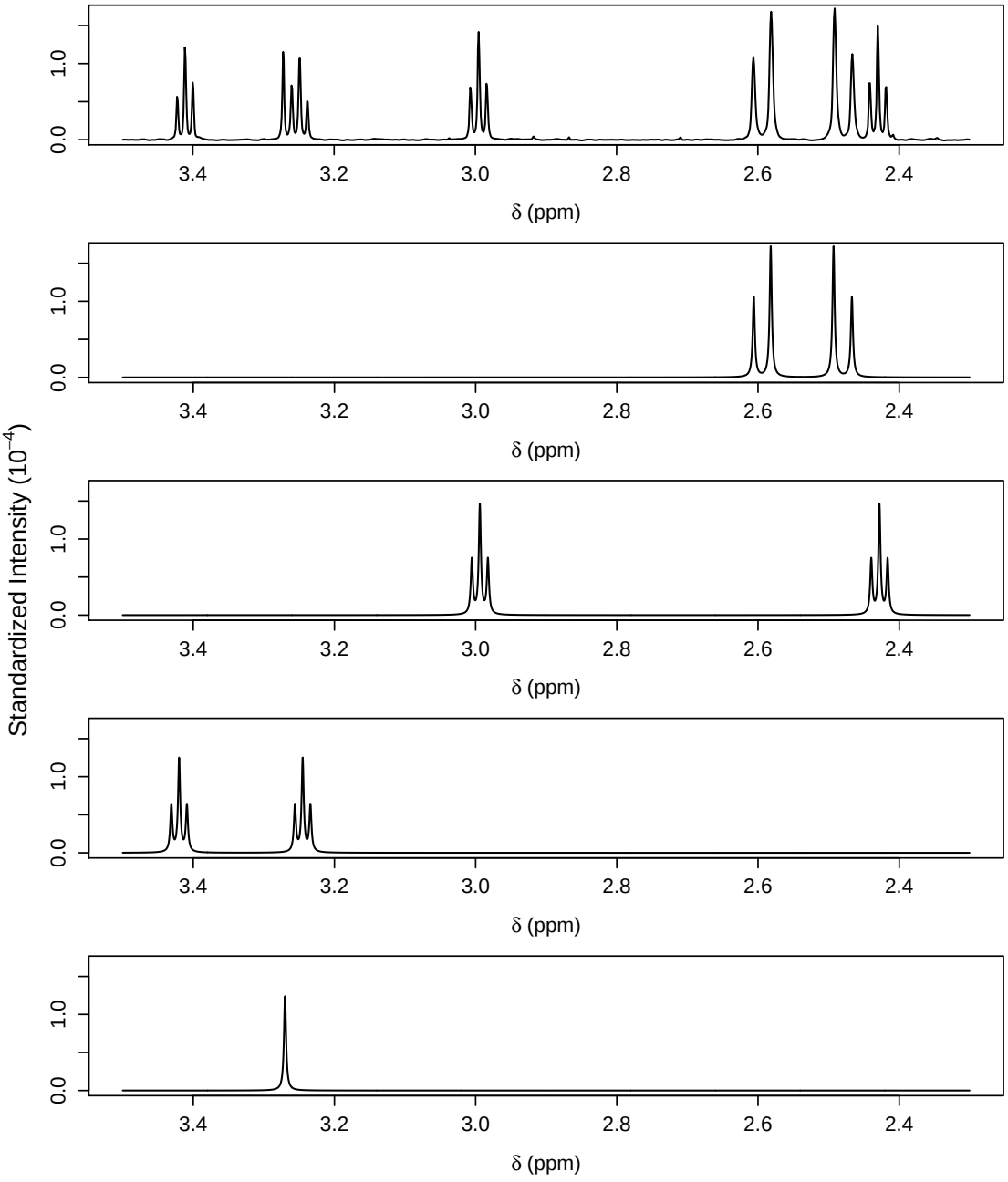


Figure 2: An ^{1}H NMR spectrum (*top panel*) with the principal resonance signals deconvolved into the metabolite NMR signatures (*lower panels*) of (in descending order by panel) citric acid, 2-oxoglutaric acid, taurine and trimethylamine N-oxide.

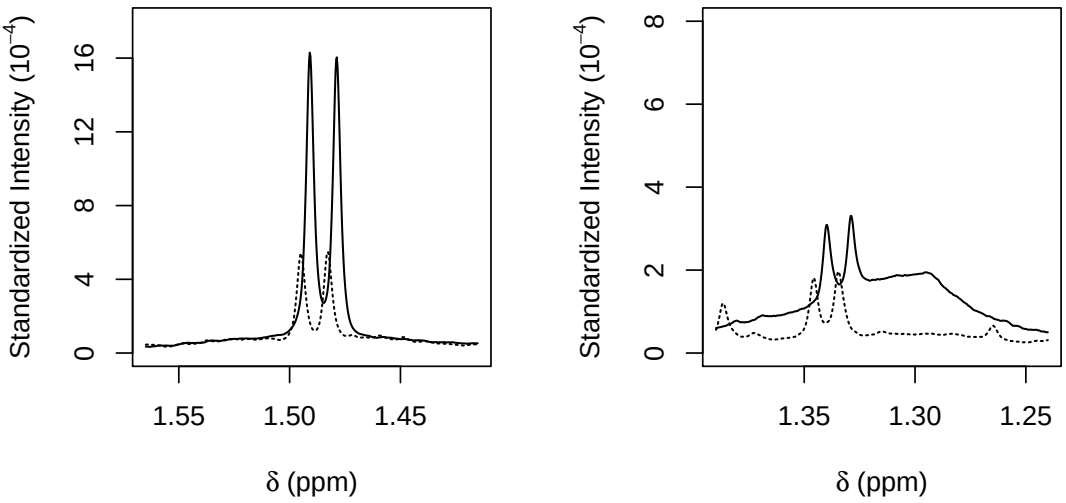


Figure 3: Positional noise between, and peak overlap within, two NMR spectra taken from the yeast experiment; resonances are generated by alanine (*left*) and threonine (*right*).

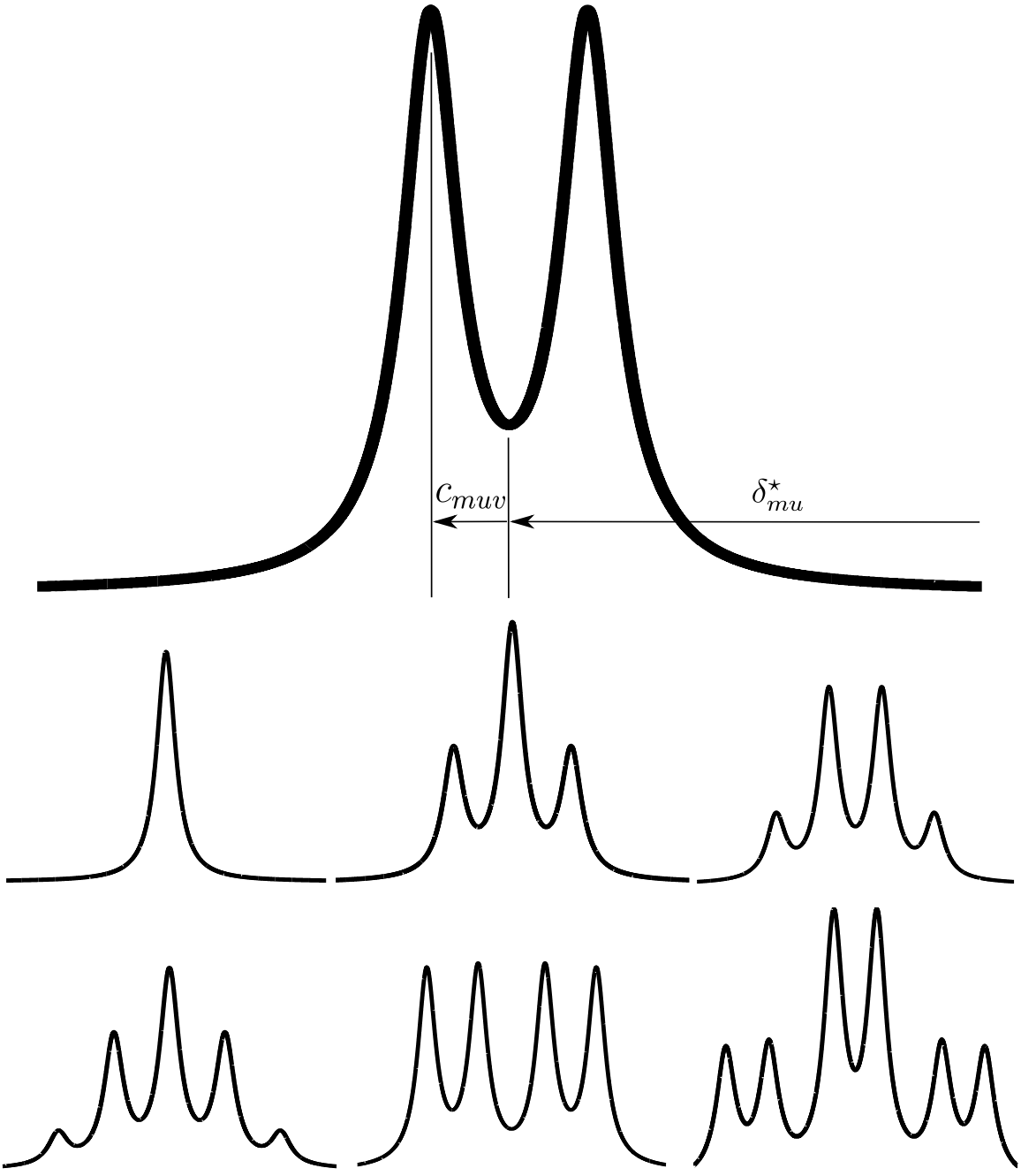


Figure 4: The peak configurations of some common types of multiplet. *Top row*: a doublet, with chemical shift $\delta^{\star}_{mu}$ and peak offset c_{muv}. *Middle row*: (from left to right) a singlet, a triplet, a quadruplet. *Bottom row*: (from left to right) a quintuplet, a doublet of doublets and a triplet of doublets.

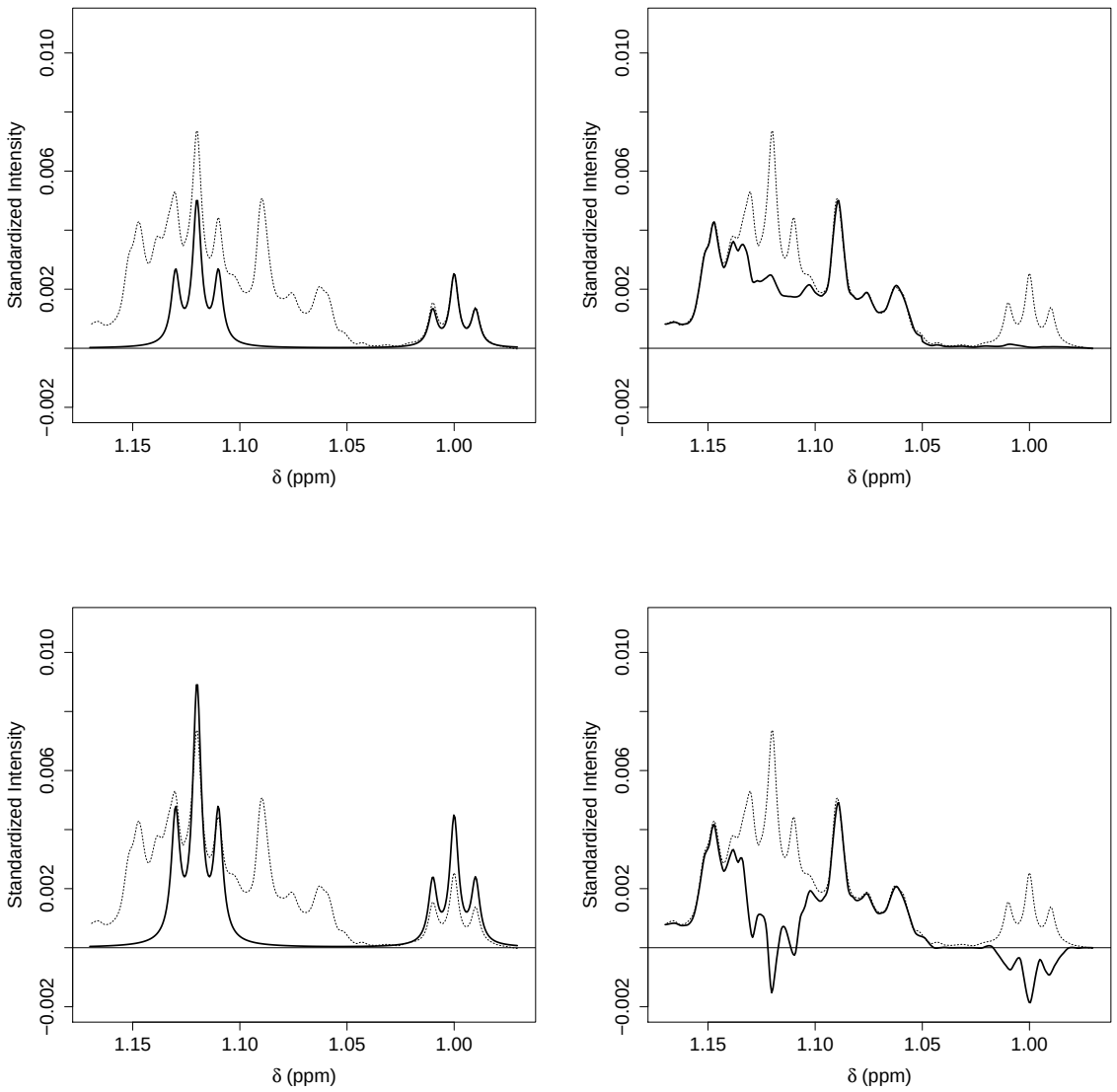


Figure 5: The effect of a prior penalizing the ξ component of the likelihood in the lower half plane (*top*) compared to one without this penalization (*bottom*). The dashed lines show the spectral data. Deconvolution of a parametric metabolite signature template (heavy lines, *left*) can be more accurate when the wavelet component (heavy lines, *right*) is penalized below the δ-axis.

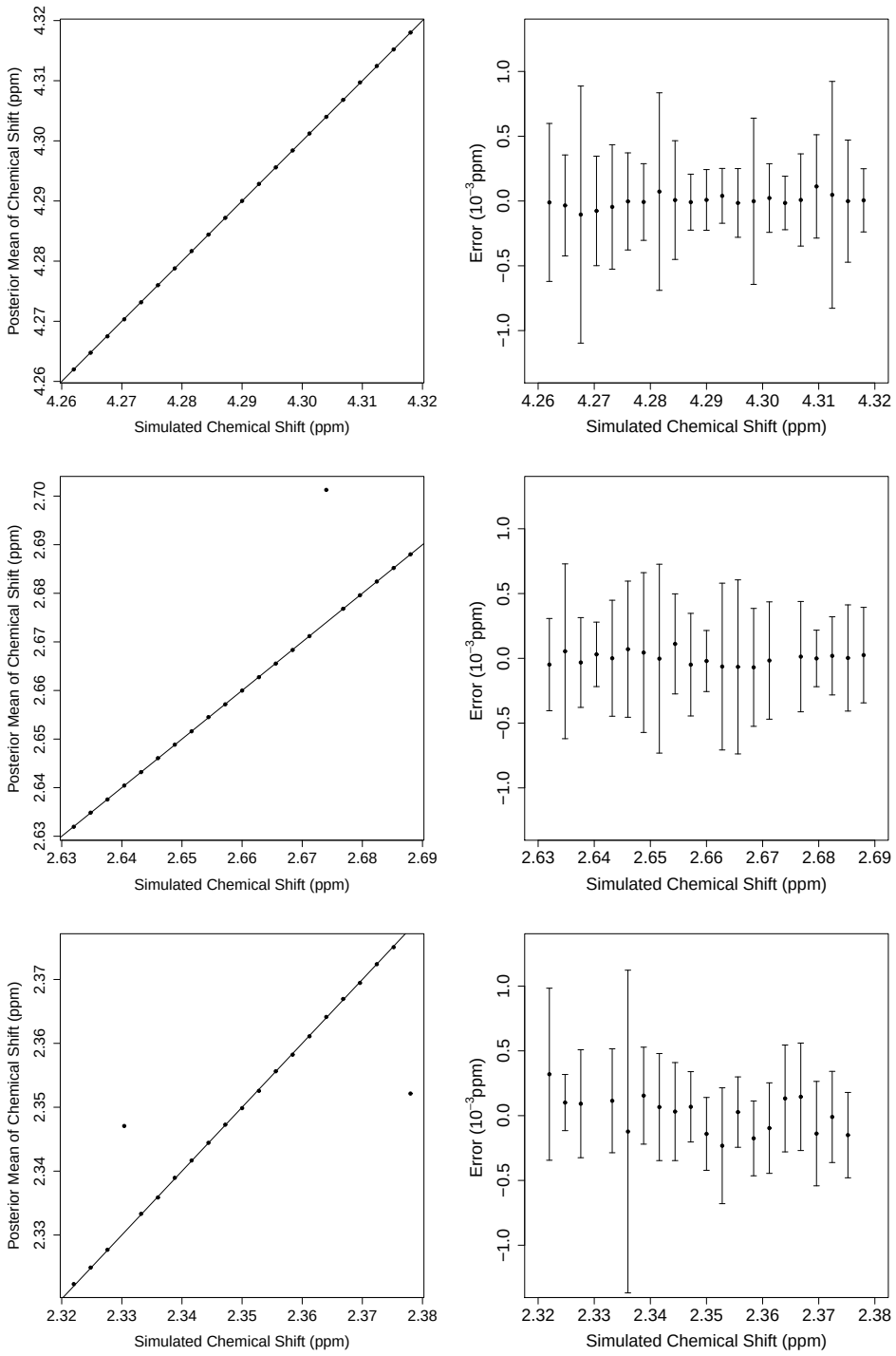


Figure 6: *Left*: posterior mean chemical shift estimates plotted against the corresponding simulated values for the three multiplets of malic acid (in rows). *Right*: estimation error, dots indicate difference between posterior mean estimates and simulated values, bands indicate 95% credible intervals. (Outliers out of plotting range.)

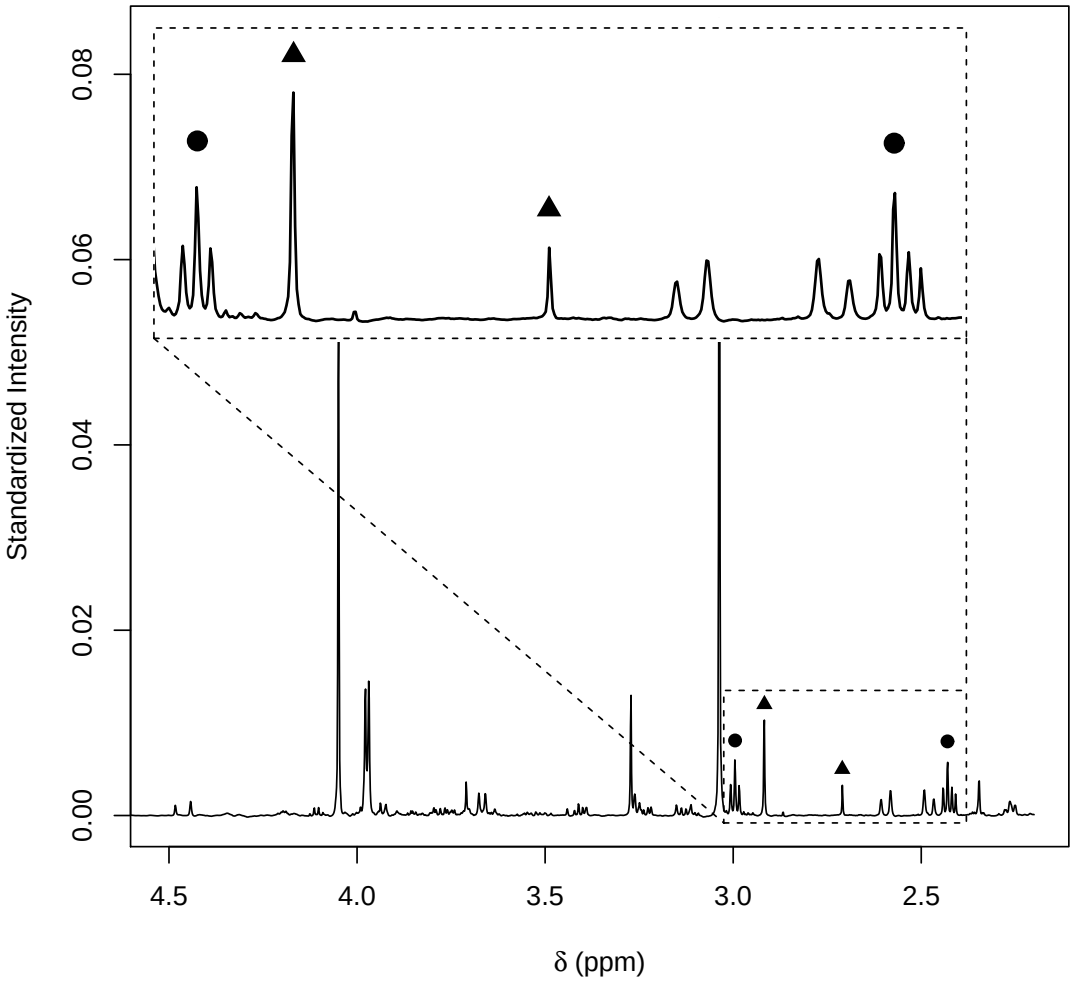


Figure 7: One of the MetAssimulo generated spectra used in the simulations to assess concentration estimation. The multiplets generated by 2-oxoglutaric acid are marked with bullets (●) while the multiplets generated by dimethylglycine are marked with triangles (▲). Note the overlap of the triplet at 2.42ppm with another uncharacterized resonance.

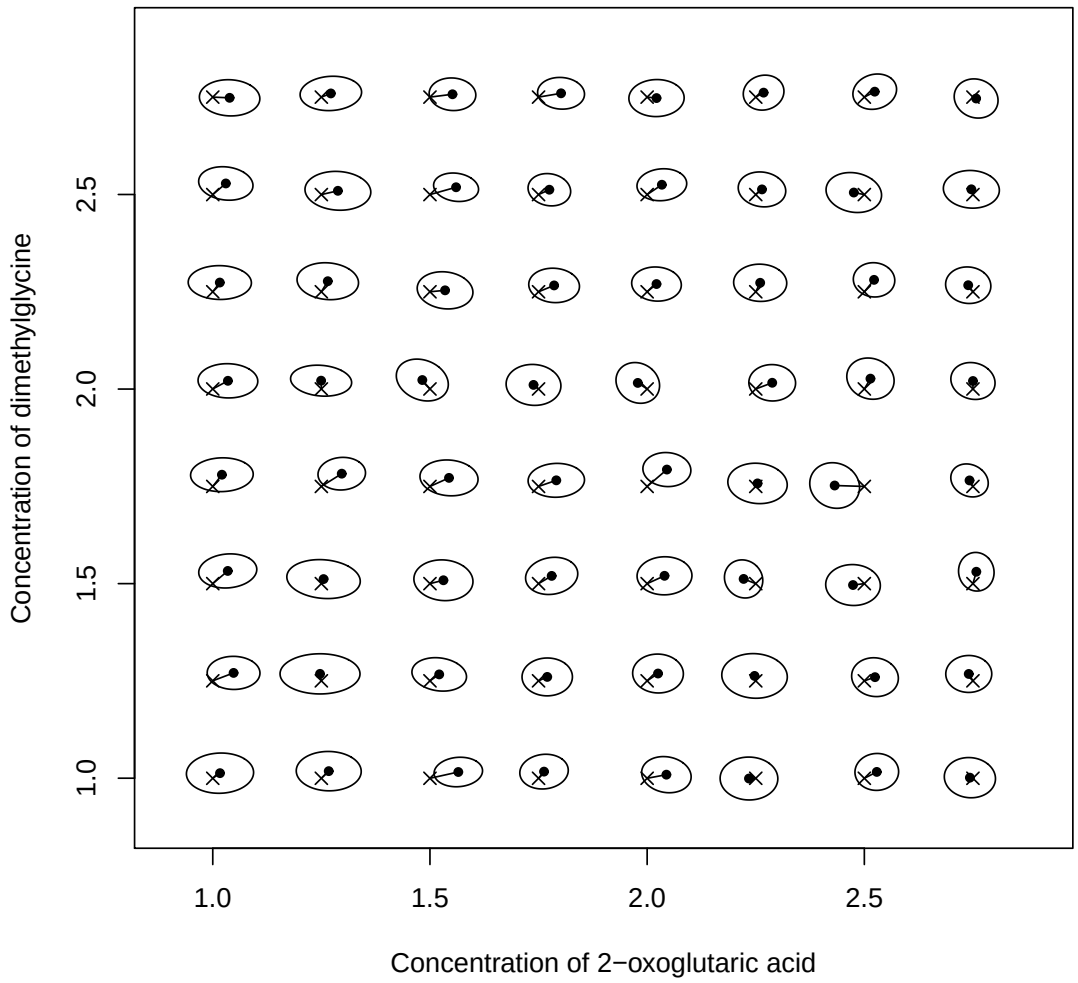


Figure 8: Inference of the concentrations of 2-oxoglutaric acid (x-axis) and dimethylglycine (y-axis). Each ellipse and dot-cross pairing corresponds to a single simulated spectrum. The crosses indicate the simulated concentrations, the black dots are Monte Carlo estimates of the posterior mean concentrations and the ellipses are Monte Carlo estimates of 95% credible regions based on bivariate Gaussian approximations. The scale has been fixed so that the smallest simulated concentration is equal to 1.0.

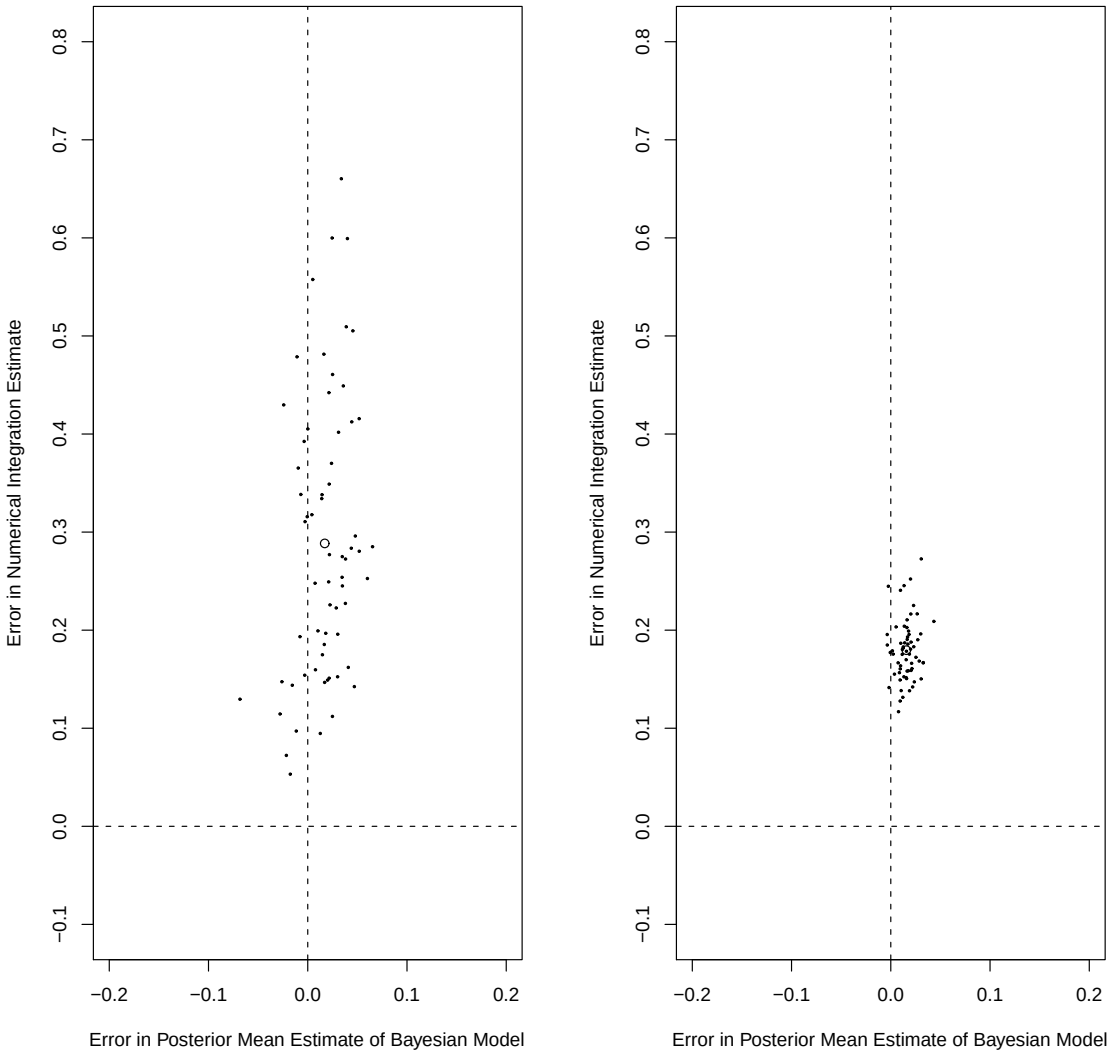


Figure 9: Bias in the estimates of concentrations of 2-oxoglutaric acid (*left*) and dimethylglycine (*right*) using the Bayesian model (x-axes) and numerical integration (y-axes), across the 64 simulation replicates. The open circles ($\circ$) show the mean estimation error for each metabolite. The scale is that of Figure 8.

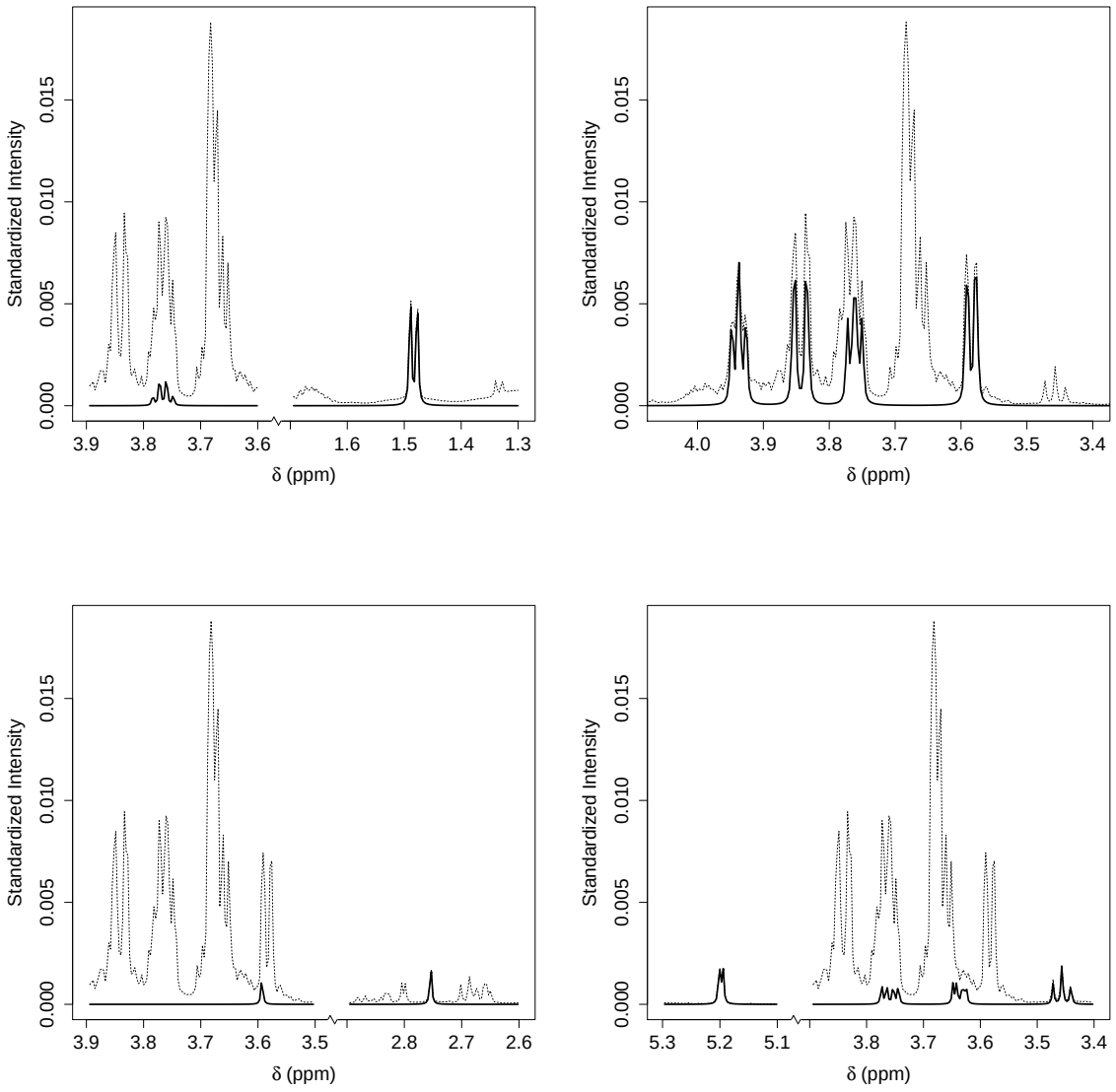


Figure 10: Deconvolution of selected metabolites from one of the yeast spectra: alanine (*top-left*), arabitol (*top-right*), trehalose (*bottom-left*) and methionine sulfoxide (*bottom-right*). The dotted lines trace the original spectral data over a suitable range. The heavy lines shows the posterior mean deconvolution (sampled on the same grid as the original spectrum), conditional on the MAP estimates of the peak-width and chemical shift parameters.

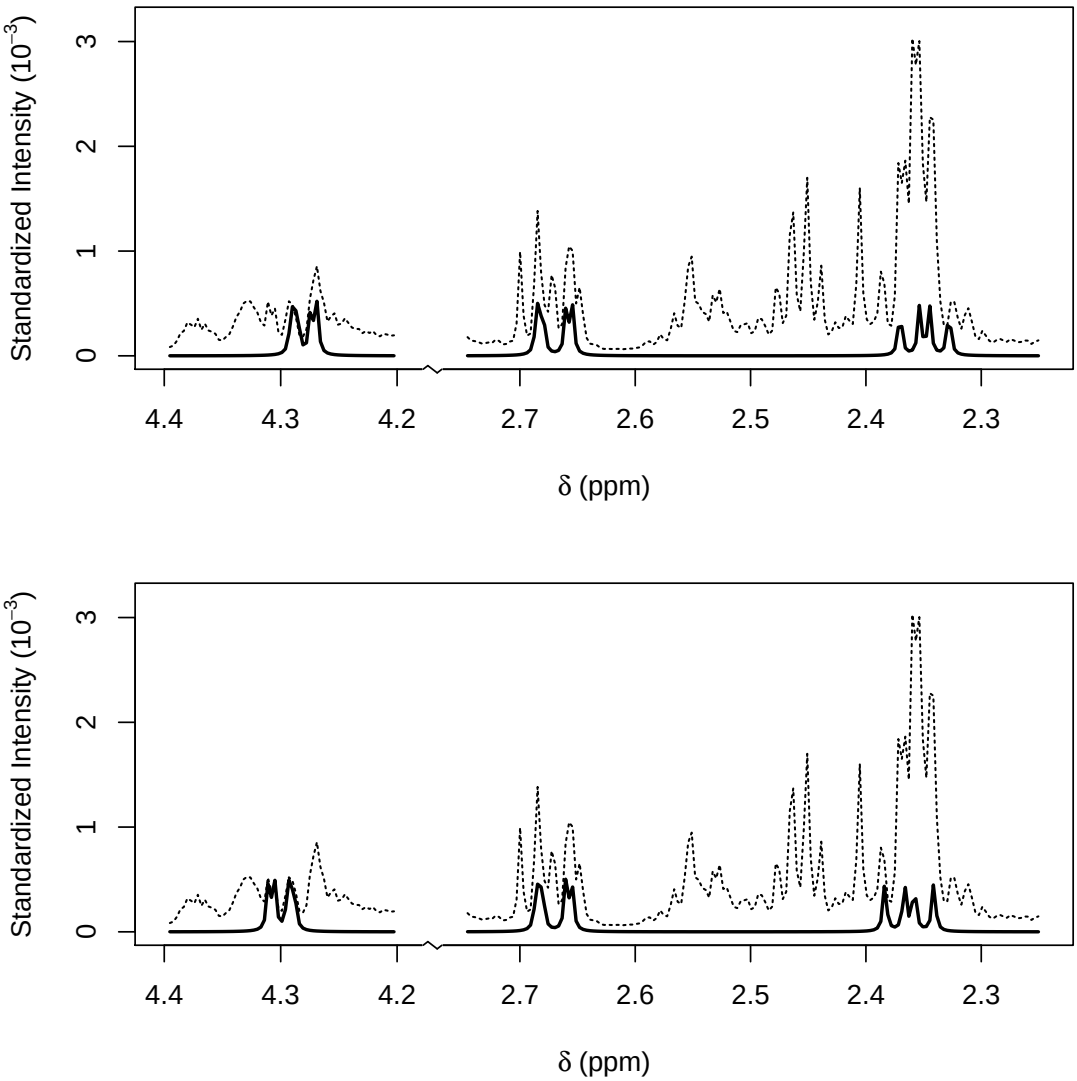
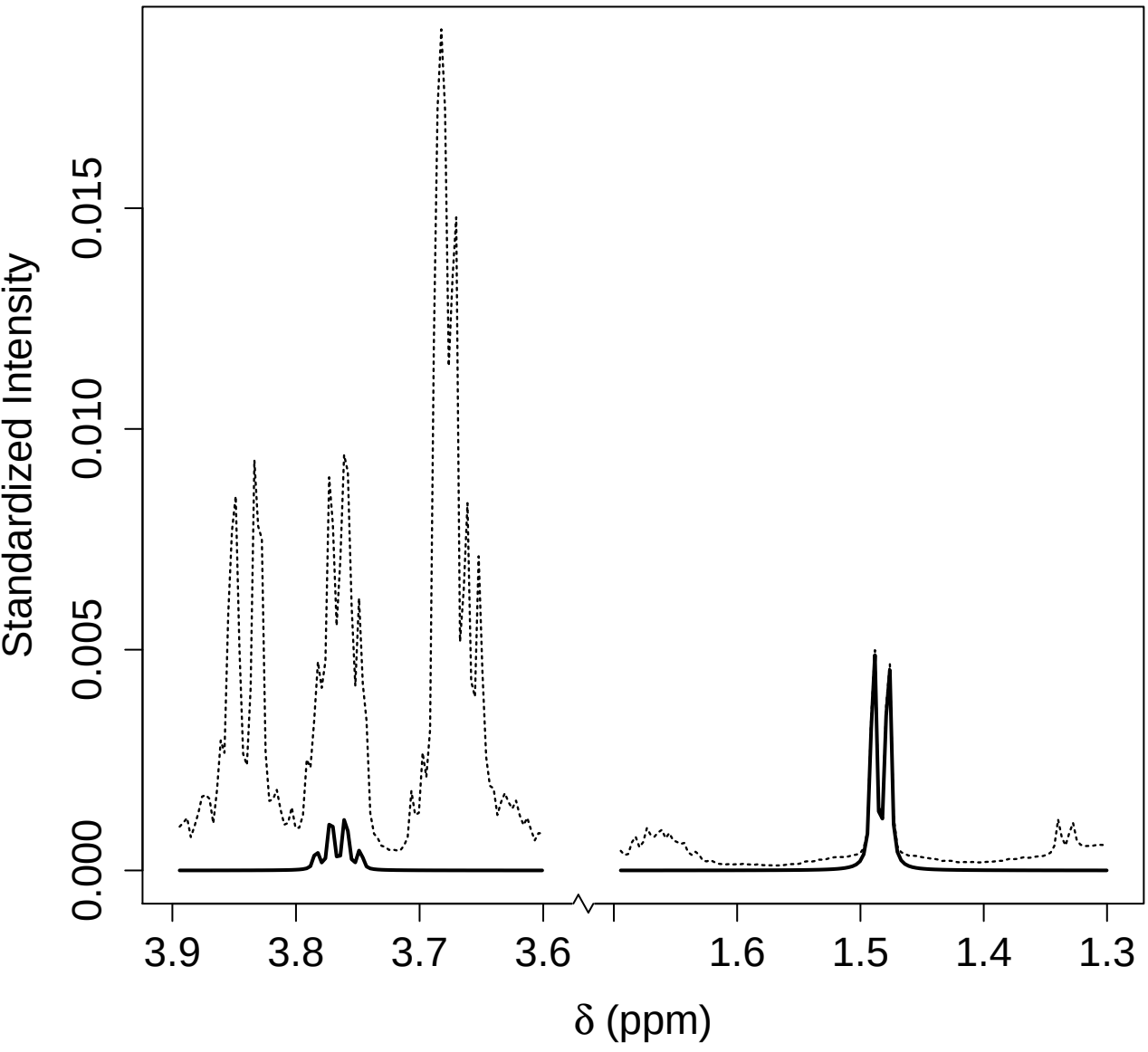
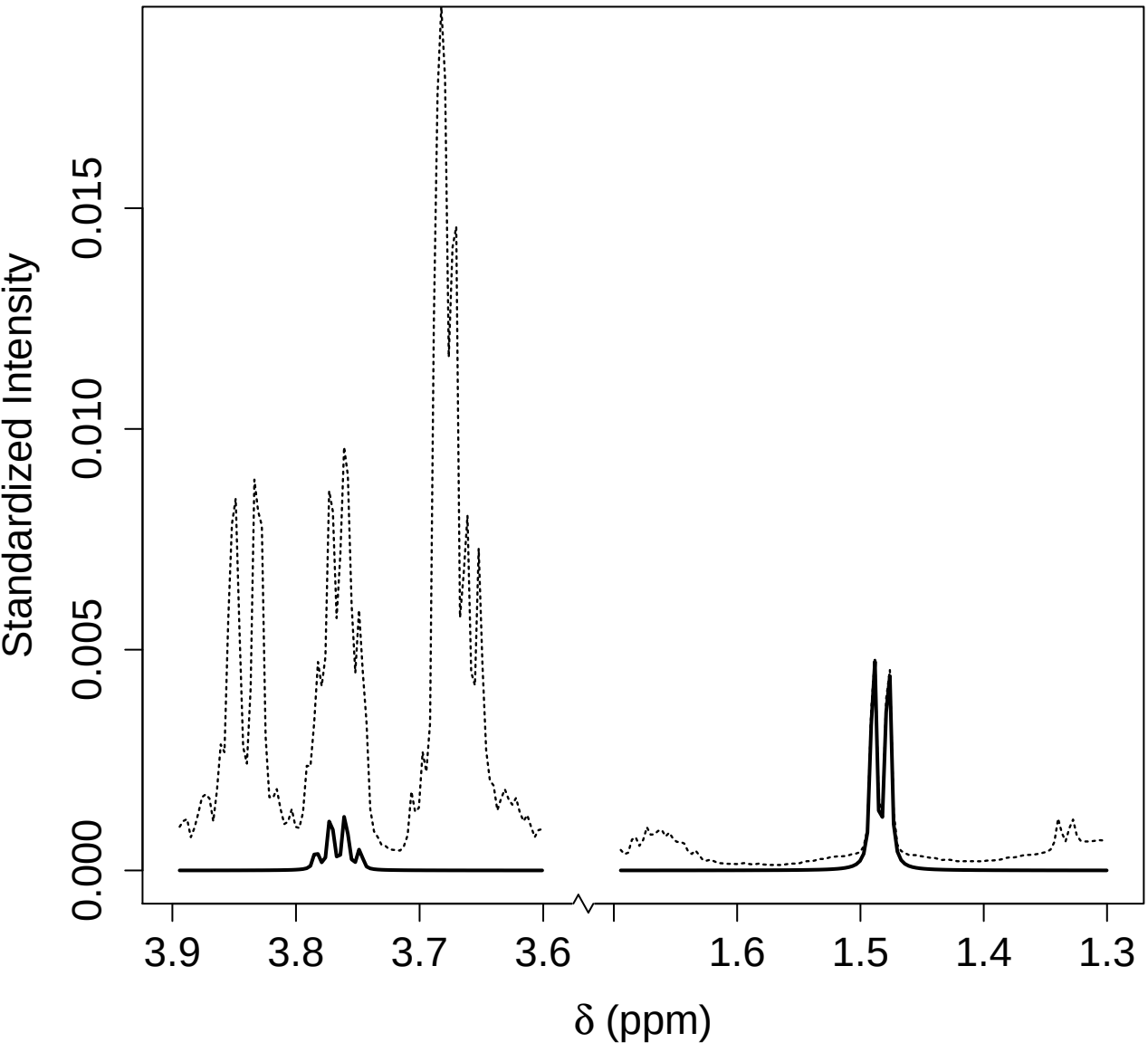
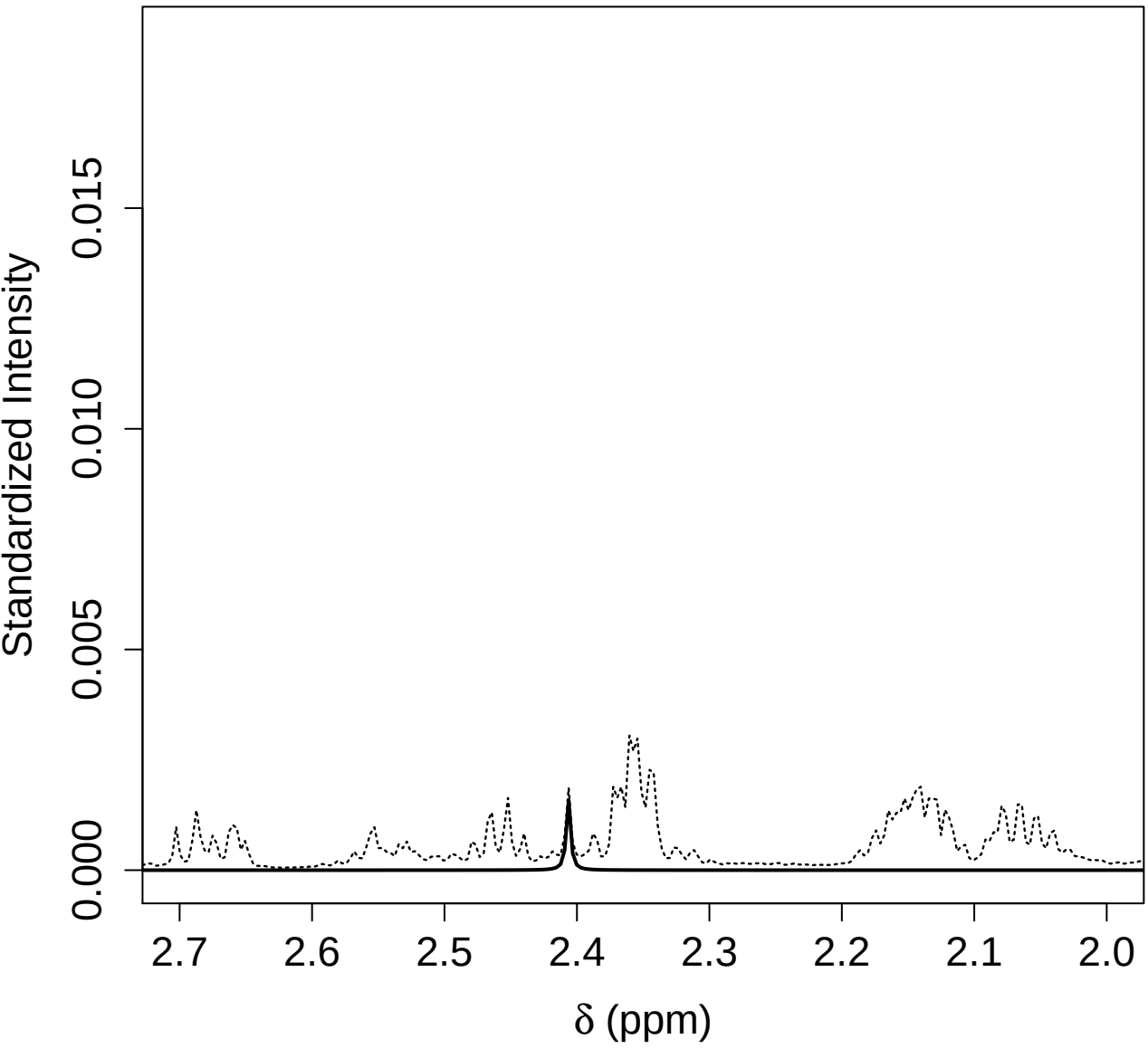
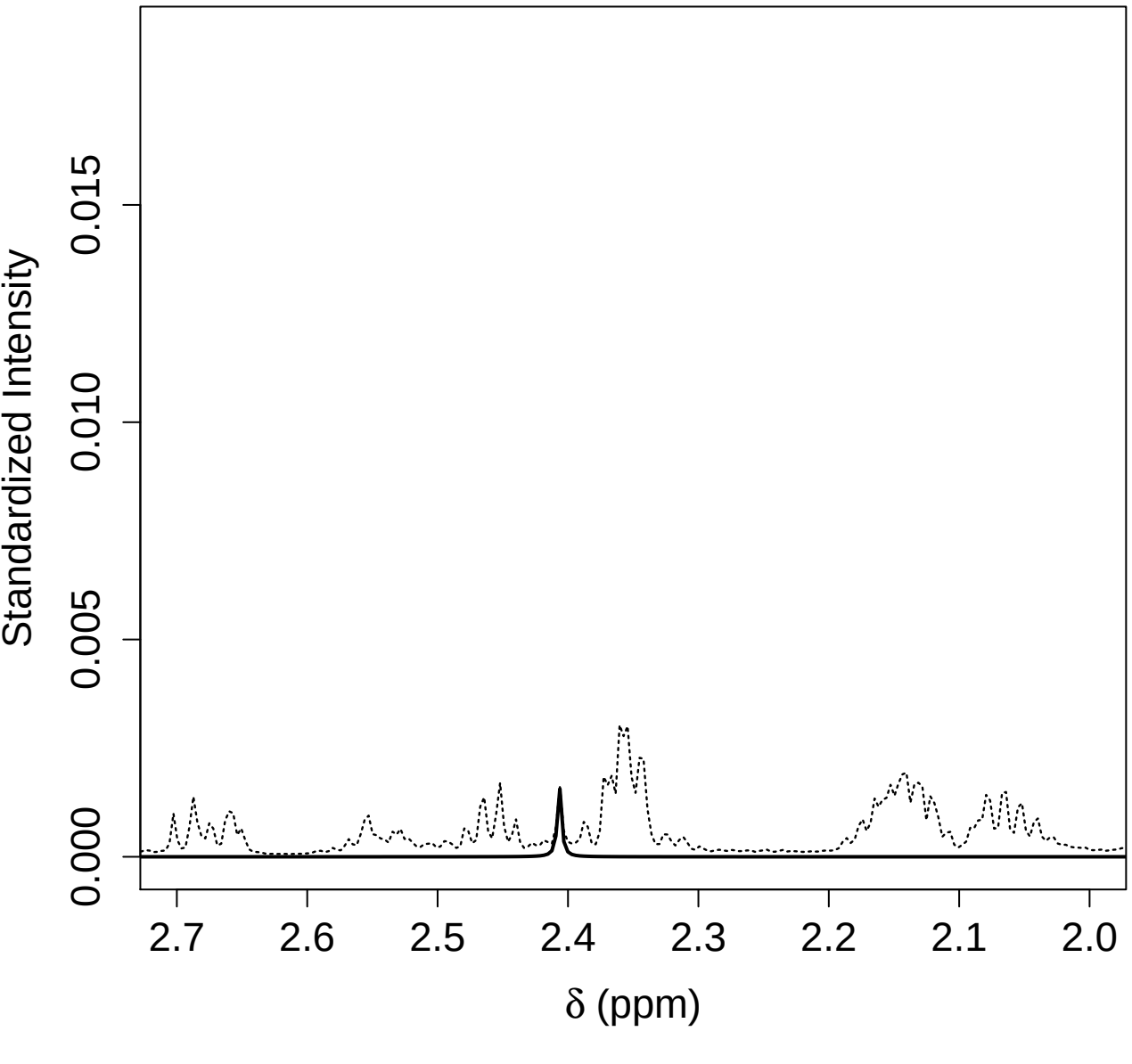


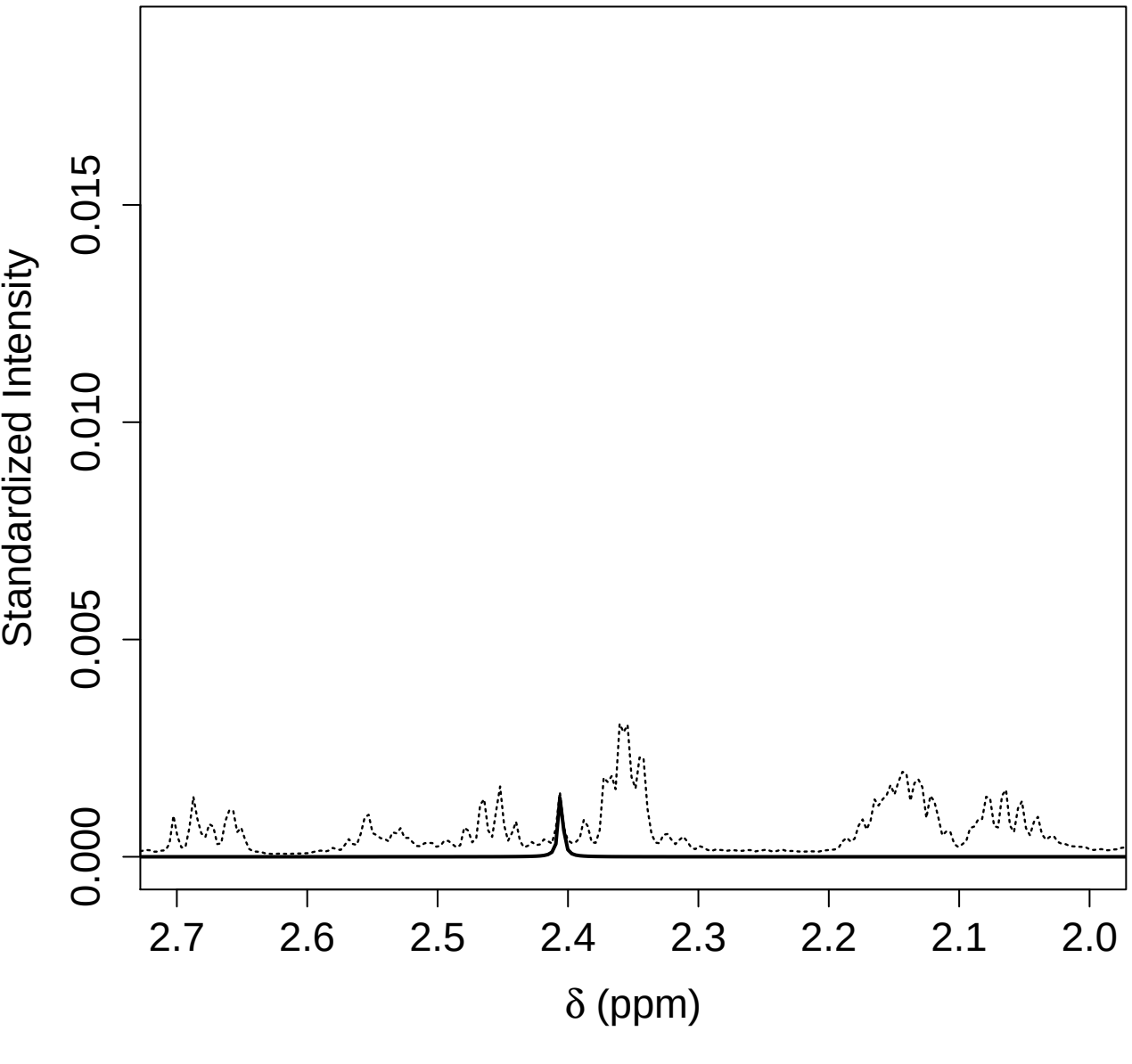
Figure 11: Two Bayesian deconvolutions of the malic acid resonances from one of the yeast spectra. *Top row*: posterior mean deconvolution with estimates of the peak-width and chemical shift parameters fixed at their MAP estimates. *Bottom row*: posterior mean deconvolution with estimates of the chemical shift parameters fixed by a spectroscopist.

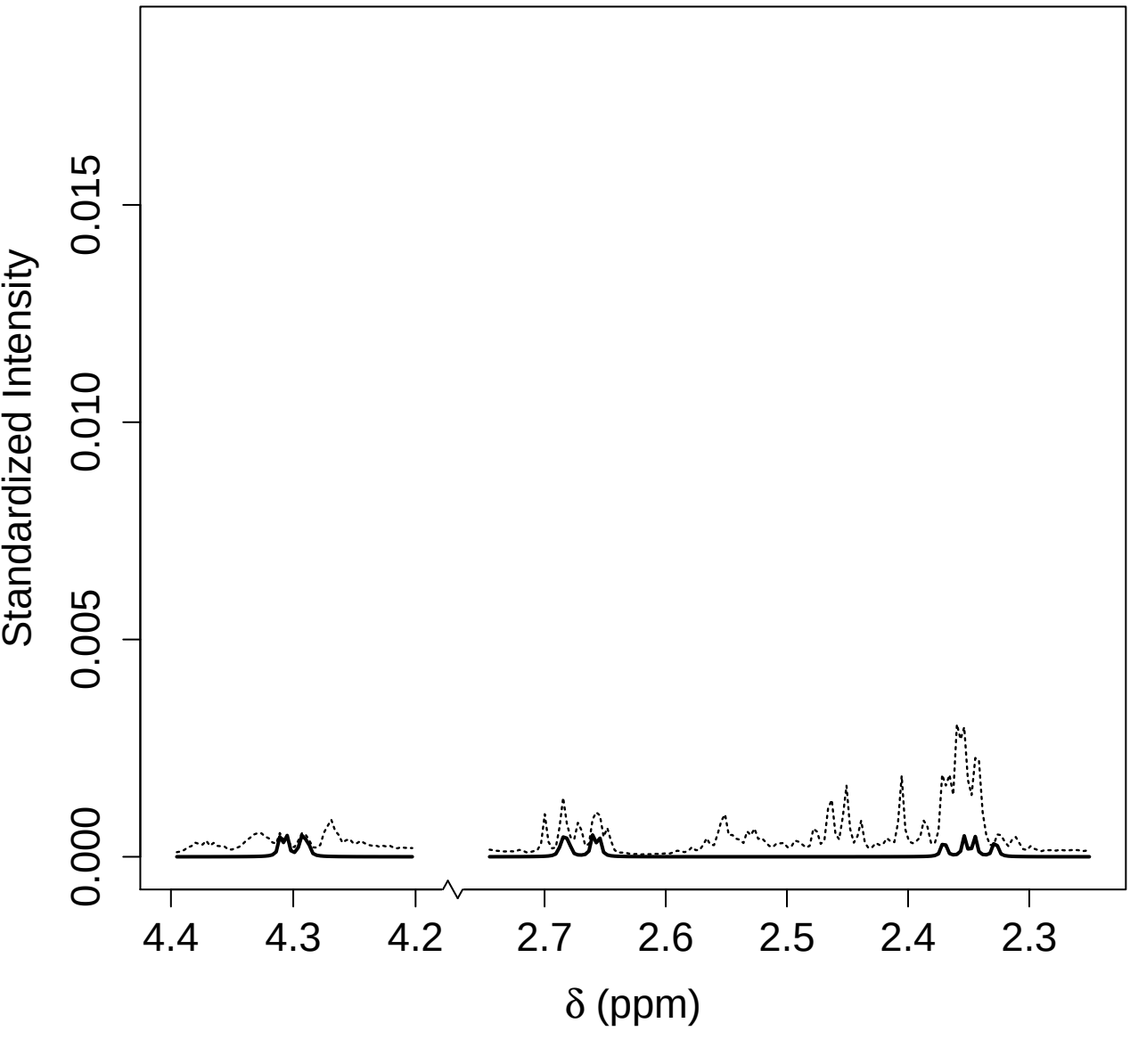

Standardized Intensity
0.000
0.005
0.010
0.015
3.9
3.8
3.7
3.6
1.6
1.5
1.4
1.3
δ (ppm)

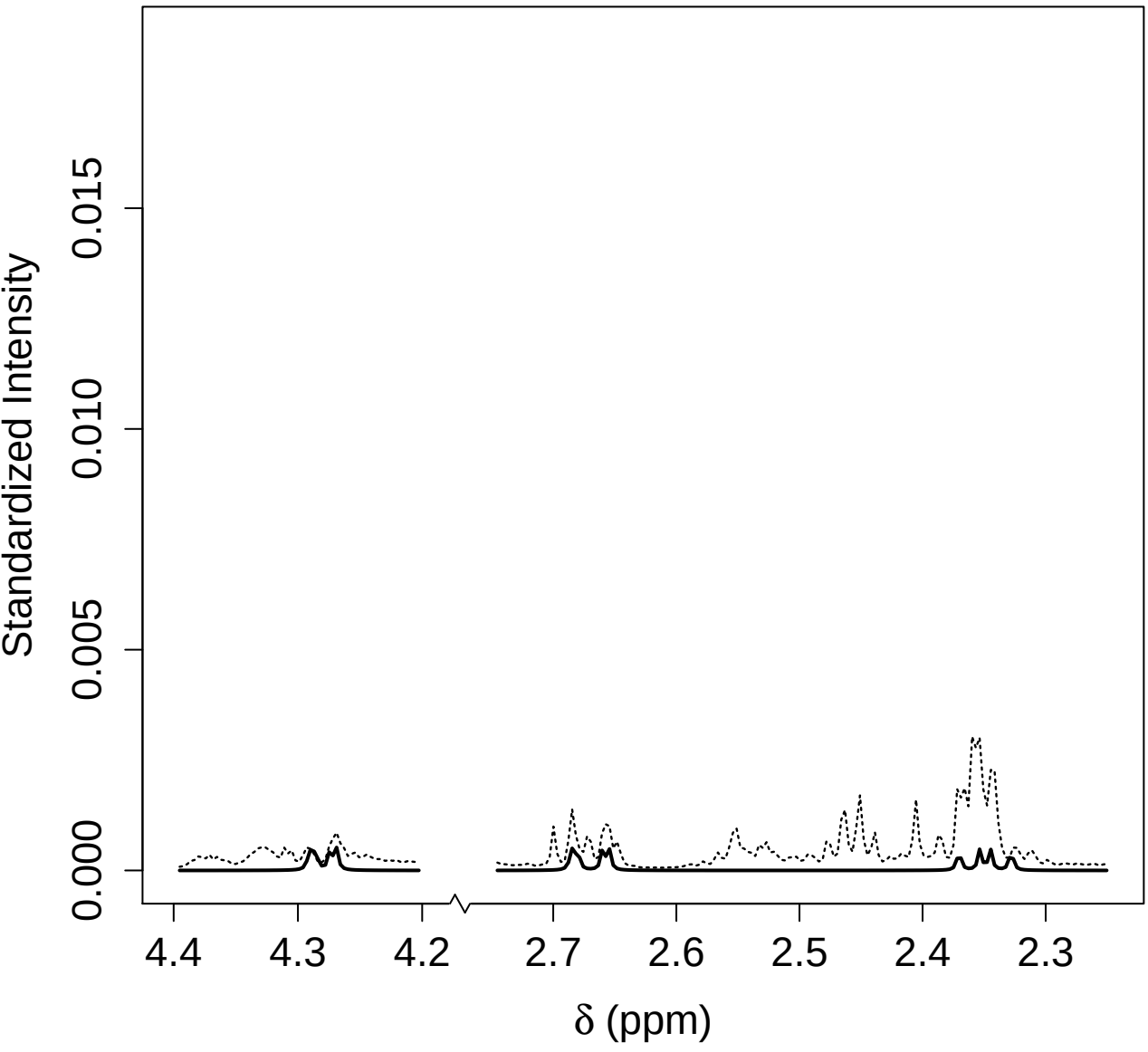

Standardized Intensity
0.000
0.005
0.010
0.015
3.9
3.8
3.7
3.6
1.6
1.5
1.4
1.3
δ (ppm)

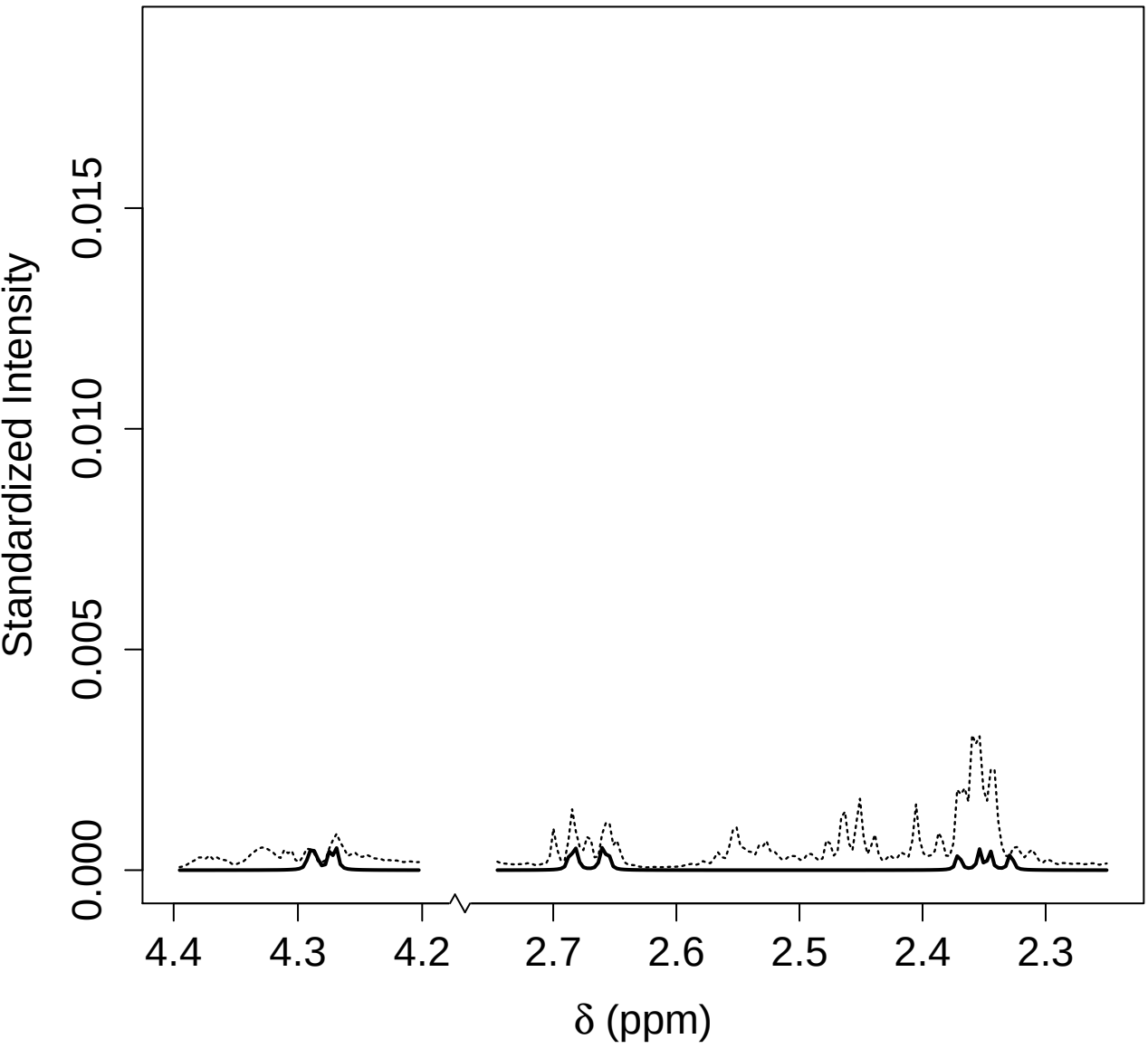

Standardized Intensity
0.000
0.005
0.010
0.015
2.7
2.6
2.5
2.4
2.3
2.2
2.1
2.0
δ (ppm)

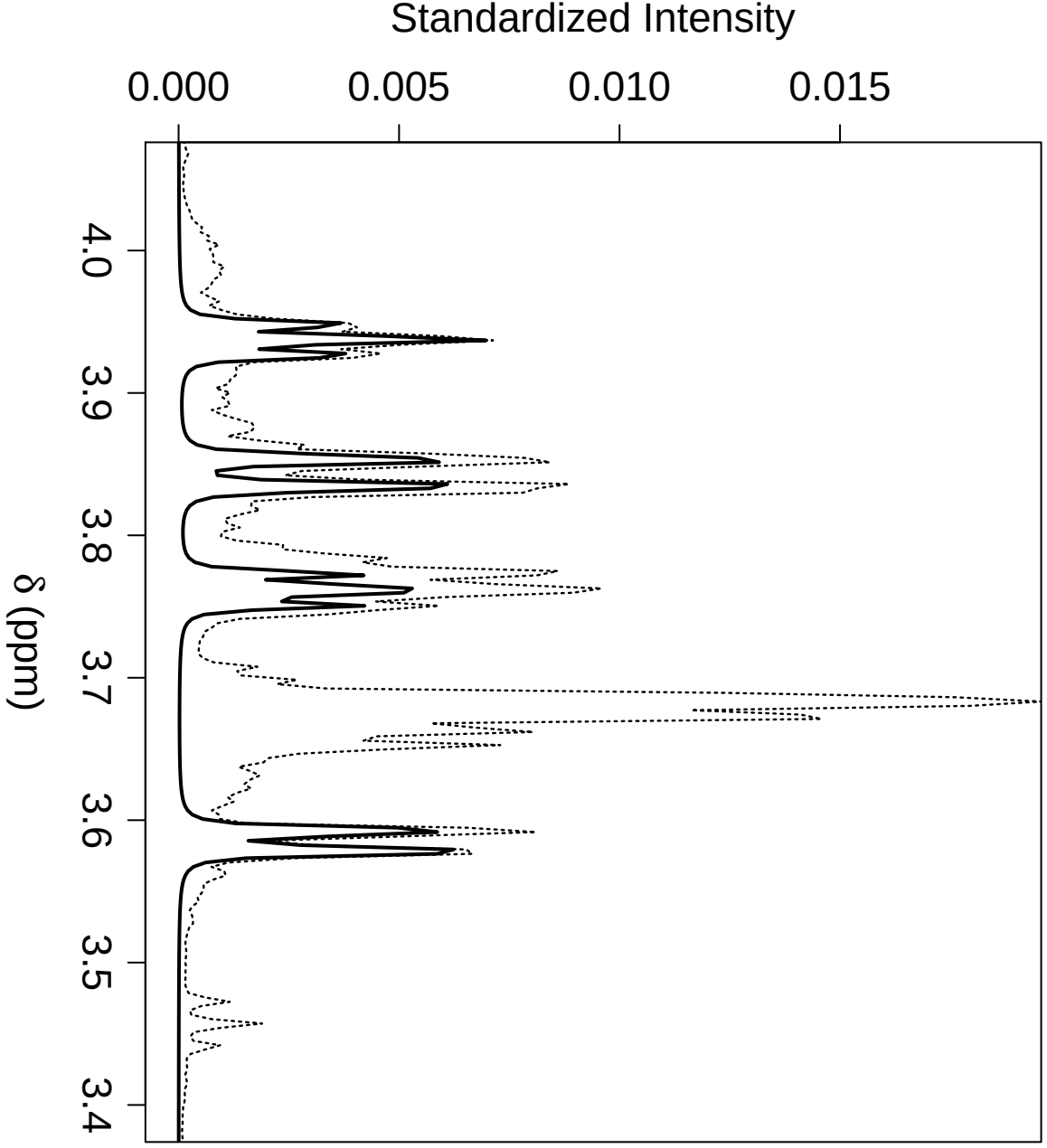

Standardized Intensity
0.000
0.005
0.010
0.015
2.7
2.6
2.5
2.4
2.3
2.2
2.1
2.0
δ (ppm)

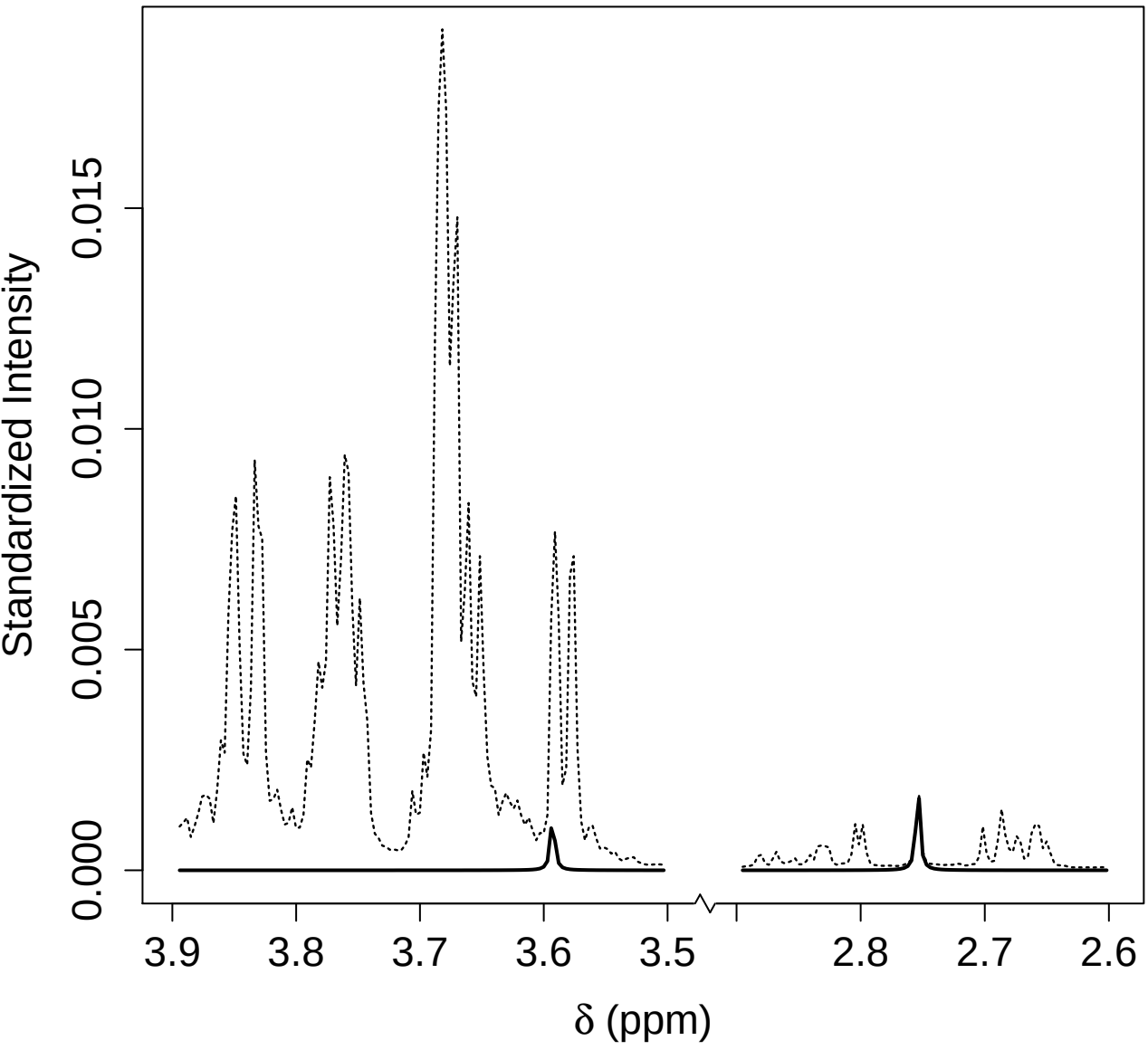

Standardized Intensity
0.000
0.005
0.010
0.015
2.7
2.6
2.5
2.4
2.3
2.2
2.1
2.0
δ (ppm)

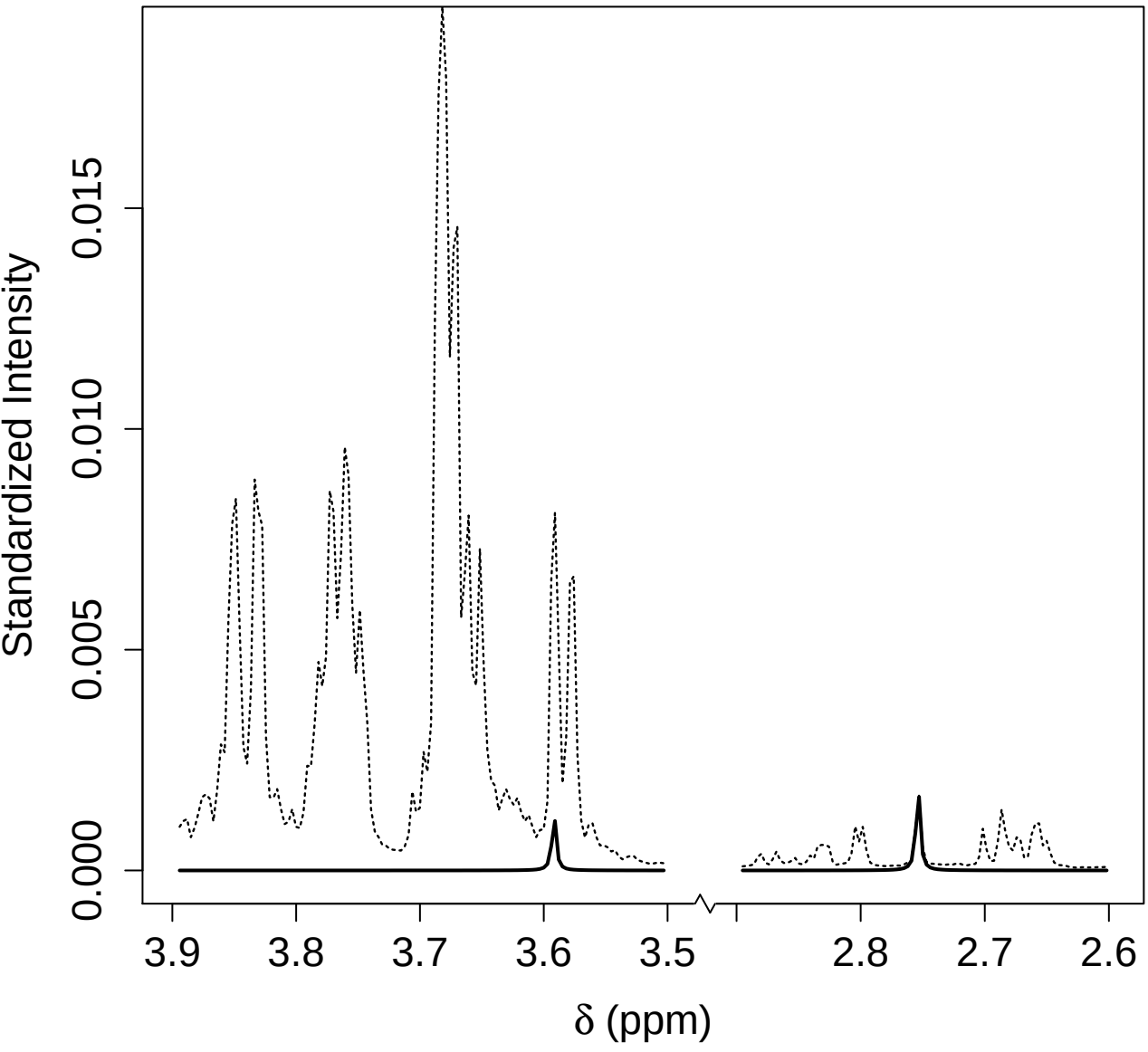

Standardized Intensity
0.000
0.005
0.010
0.015
4.4
4.3
4.2
2.7
2.6
2.5
2.4
2.3
δ (ppm)

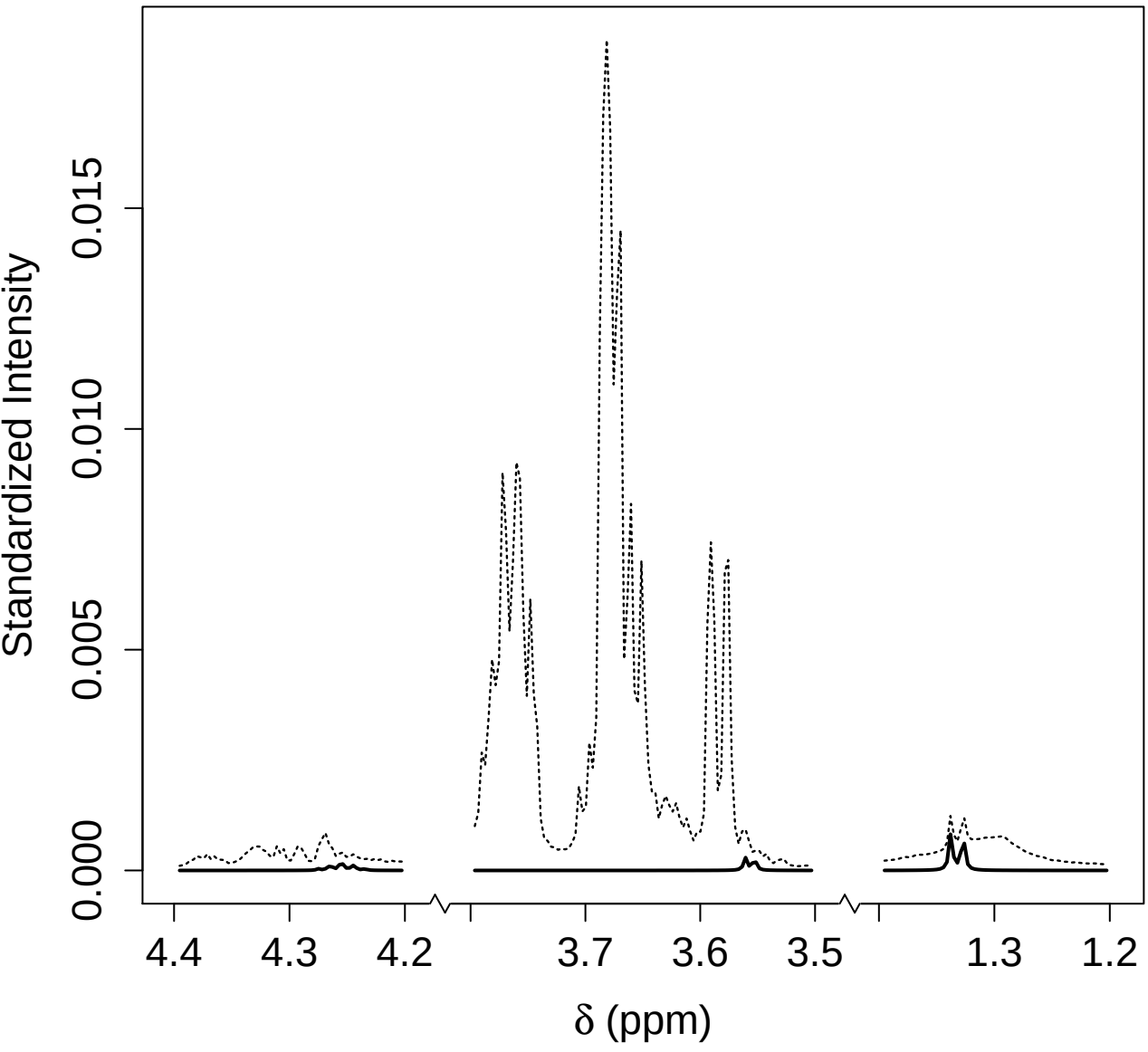

Standardized Intensity
0.000
0.005
0.010
0.015
4.4
4.3
4.2
2.7
2.6
2.5
2.4
2.3
δ (ppm)

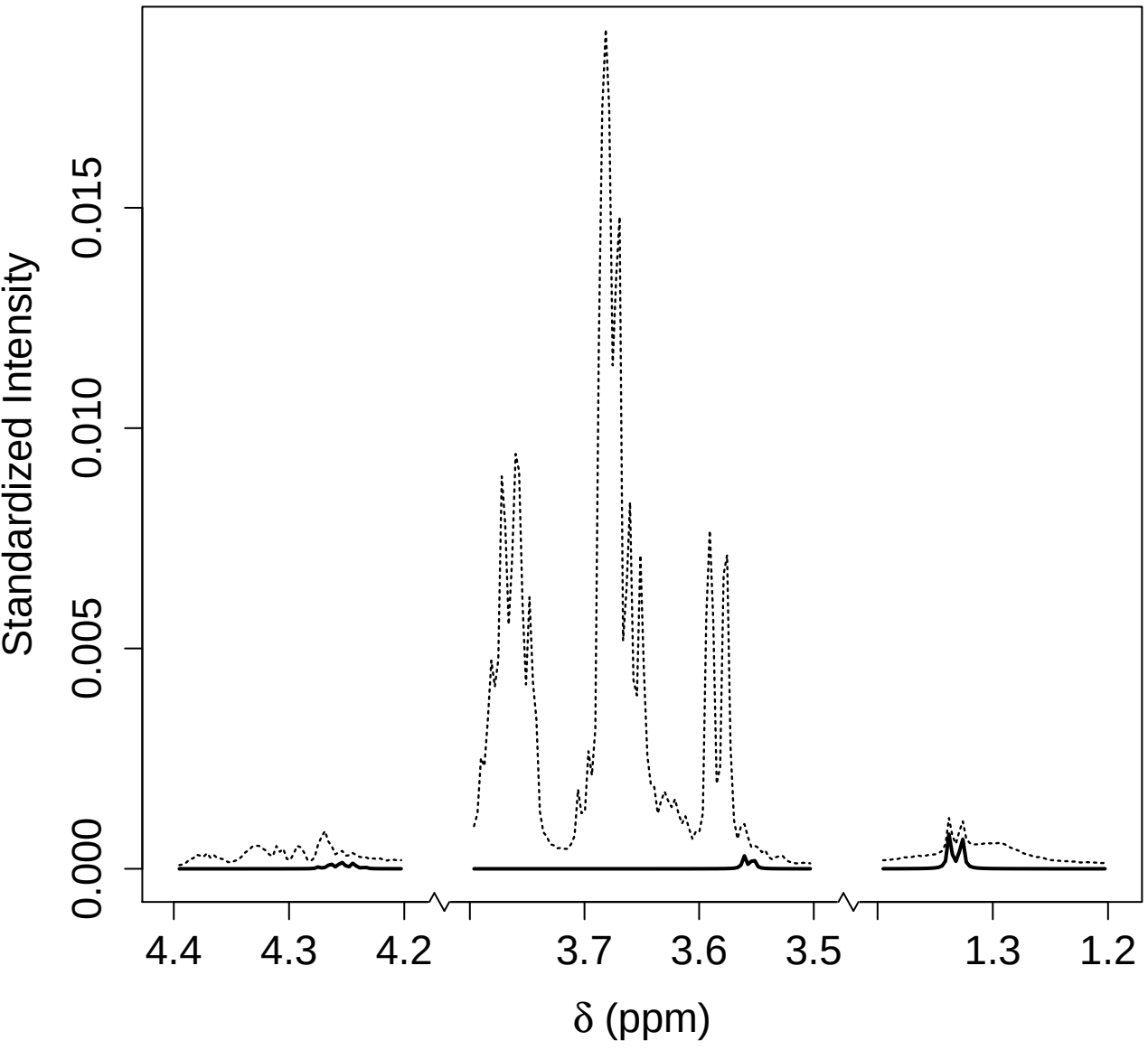

Standardized Intensity
0.000
0.005
0.010
0.015
4.4
4.3
4.2
2.7
2.6
2.5
2.4
2.3
δ (ppm)

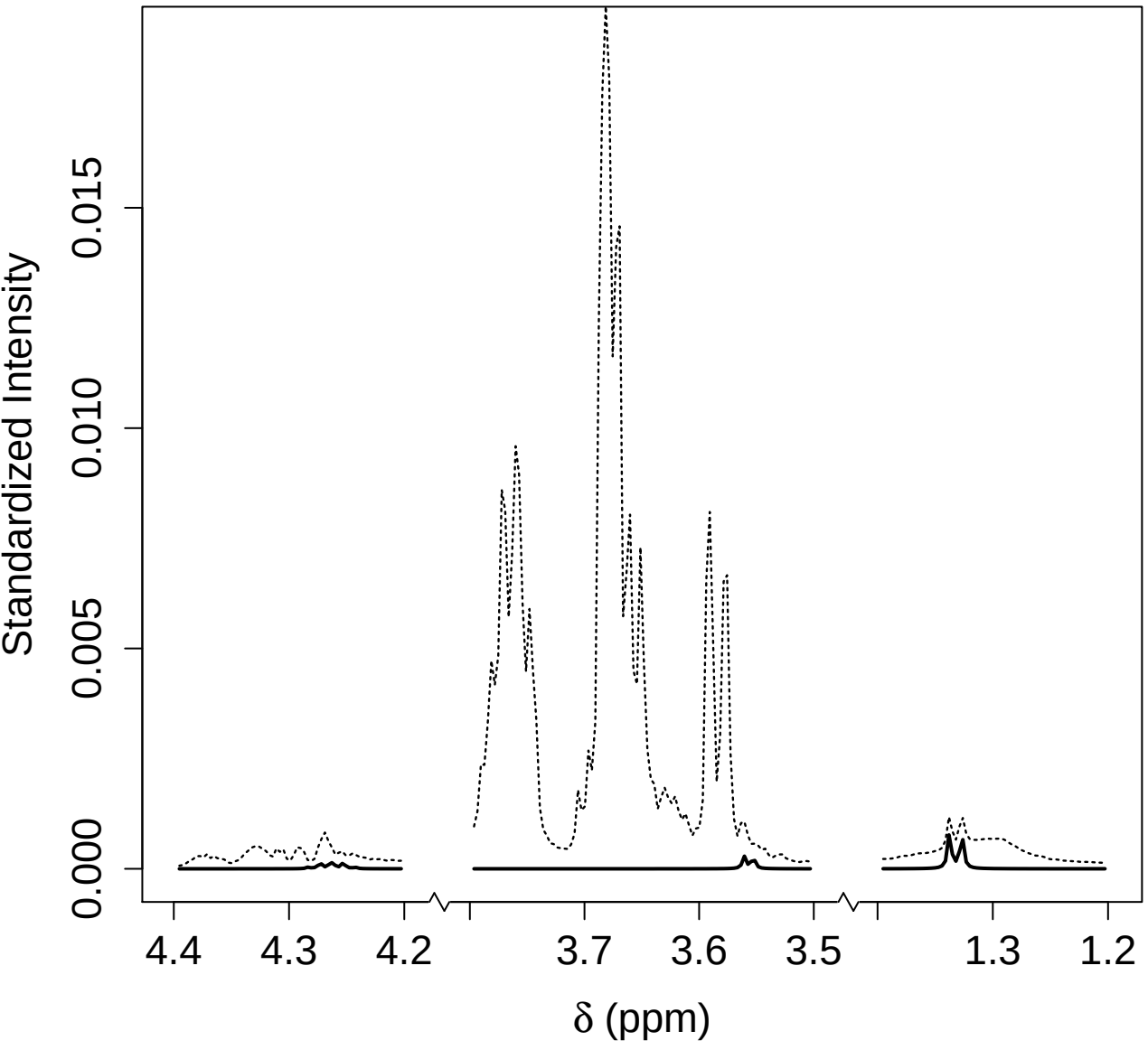

Standardized Intensity
0.000
0.005
0.010
0.015
δ (ppm)
4.0
3.9
3.8
3.7
3.6
3.5
3.4

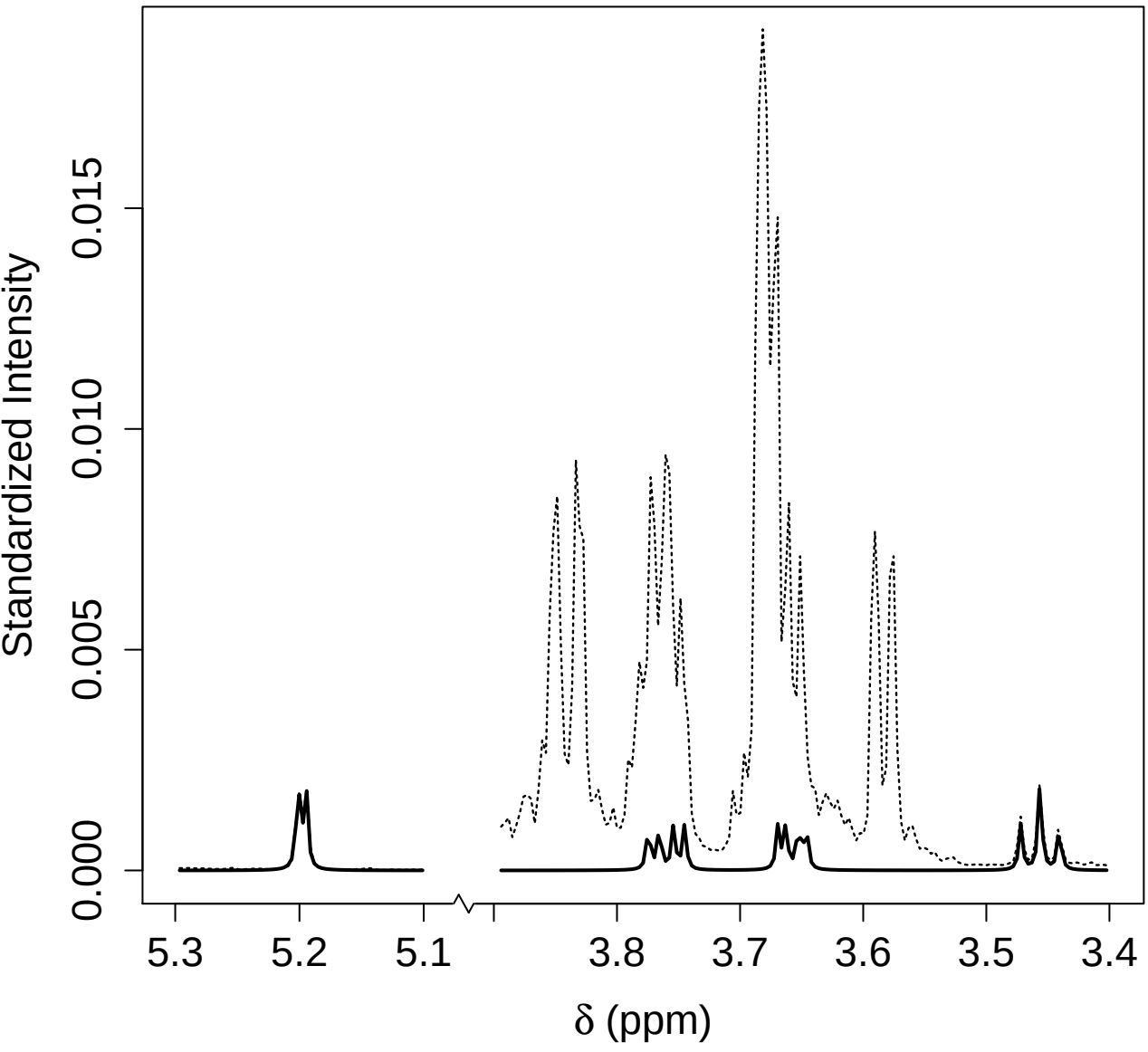

Standardized Intensity
0.000
0.005
0.010
0.015
3.9
3.8
3.7
3.6
3.5
2.8
2.7
2.6
δ (ppm)

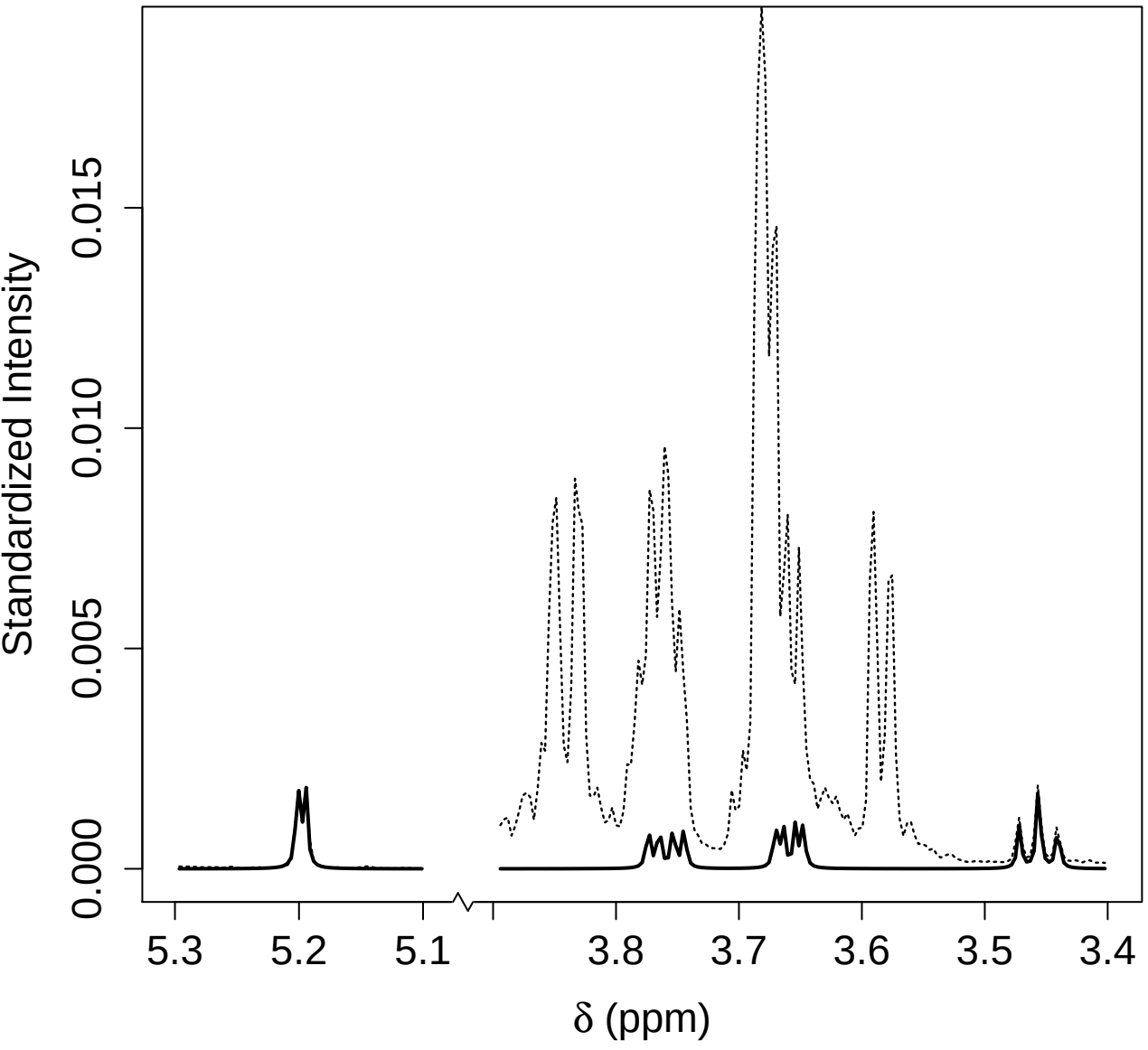

Standardized Intensity
0.000
0.005
0.010
0.015
3.9
3.8
3.7
3.6
3.5
2.8
2.7
2.6
δ (ppm)


Standardized Intensity
0.000
0.005
0.010
0.015
4.4
4.3
4.2
3.7
3.6
3.5
1.3
1.2
δ (ppm)


Standardized Intensity
0.000
0.005
0.010
0.015
4.4
4.3
4.2
3.7
3.6
3.5
1.3
1.2
δ (ppm)


Standardized Intensity
0.000
0.005
0.010
0.015
4.4
4.3
4.2
3.7
3.6
3.5
1.3
1.2
δ (ppm)


Standardized Intensity
0.000
0.005
0.010
0.015
5.3
5.2
5.1
3.8
3.7
3.6
3.5
3.4
δ (ppm)


Standardized Intensity
0.000
0.005
0.010
0.015
5.3
5.2
5.1
3.8
3.7
3.6
3.5
3.4
δ (ppm)