BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 12 2011, pages 1637–1644
doi:10.1093/bioinformatics/btr118

Gene expression

Advance Access publication March 12, 2011

Identification and quantification of metabolites in ^{1}H NMR spectra by Bayesian model selection

Cheng Zheng[1,*], Shucha Zhang[2], Susanne Ragg[3], Daniel Raftery[4] and Olga Vitek[1,5,*]

[1]Department of Statistics, Purdue University, West Lafayette, IN 47907, [2]Fred Hutchinson Cancer Research Center, Seattle, WA 98109, [3]School of Medicine, Department of Pediatrics, Indiana University, Indianapolis, IN 46202, [4]Department of Chemistry, Purdue University, West Lafayette, IN 47907 and [5]Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Nuclear magnetic resonance (NMR) spectroscopy is widely used for high-throughput characterization of metabolites in complex biological mixtures. However, accurate interpretation of the spectra in terms of identities and abundances of metabolites can be challenging, in particular in crowded regions with heavy peak overlap. Although a number of computational approaches for this task have recently been proposed, they are not entirely satisfactory in either accuracy or extent of automation.

Results: We introduce a probabilistic approach Bayesian Quantification (*BQuant*), for fully automated database-based identification and quantification of metabolites in local regions of ^{1}H NMR spectra. The approach represents the spectra as mixtures of reference profiles from a database, and infers the identities and the abundances of metabolites by Bayesian model selection. We show using a simulated dataset, a spike-in experiment and a metabolomic investigation of plasma samples that *BQuant* outperforms the available automated alternatives in accuracy for both identification and quantification.

Availability: The R package *BQuant* is available at: http://www.stat.purdue.edu/~ovitek/BQuant-Web/.

Contact: ovitek@stat.purdue.edu; zhengc@purdue.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on December 6, 2010; revised on February 20, 2011; accepted on February 25, 2011

1 INTRODUCTION

The field of *metabolomics* studies small molecules such as amino acids, nucleic acids, lipids and carbohydrates, called *metabolites*, which are present in cells and extracellular fluids of biological organisms (Wishart *et al.*, 2007). Metabolites are the end products of a variety of cellular processes and their high-throughput characterization and quantification can provide important insights into the functioning of living organisms. As a result, metabolomics is increasingly applied to the areas of system biology, drug discovery, pharmaceutical research, early disease detection, toxicology and food science (Gowda *et al.*, 2008).

*To whom correspondence should be addressed.

Developments in high-throughput metabolomics are driven by advances in analytical methods (Chen *et al.*, 2006), such as nuclear magnetic resonance (NMR) spectroscopy and in mass spectrometry. In particular, ^{1}H NMR spectroscopy is non-destructive and highly reproducible, and is widely used for the identification and quantification of metabolites in biofluids such as plasma and urine (Nicholson *et al.*, 1999). The ^{1}H NMR signal is generated by the motion of magnetic moments of protons or other nuclei in a magnetic field after their excitement with a high-frequency pulse (Silverstein *et al.*, 2005). Fourier transformation of this time-dependent signal yields *a chemical shift*, a measure that expresses the dependence of nuclear magnetic energy on the electronic and chemical environment in the molecule. An ^{1}H NMR spectrum displays chemical shifts in parts per million (p.p.m., i.e. the difference in hertz between a resonance frequency and that of a reference substance, over frequency of magnet field in megahertz), versus intensities of the signals (Ernst *et al.*, 1990).

Typically, a metabolite detectable by ^{1}H NMR contains one or more protons, and each of the protons produces one or more peaks. The number of peaks generated by a metabolite, as well as their location and ratio of heights, are reproducible and uniquely determined by the chemical structure of the molecule. Figure 1 illustrates two ^{1}H NMR spectra with a dominant signal from metabolite taurine from the spike-in experiment in Section 4. All three predominant peaks in the spectra correspond to one proton in the taurine chemical structure. The expected ratio of heights of the three peaks is known and fixed to 1:2:1.

Extensive information regarding NMR-specific characteristics of a variety of metabolites is currently available. Several public databases, such as human metabolome database (HMDB) (Wishart *et al.*, 2007) and BioMagResBank (BMRB) (Ulrich *et al.*, 2008), as well as commercial databases such as the one in the Chenomx NMR suite (Weljie *et al.*, 2006), store reference NMR spectra of metabolites. The reference spectra can contain observed peak locations and ratios of heights of spectra from pure chemical compounds, or their simulated counterparts. Therefore, in principle, researchers can identify metabolites by matching the observed spectra to the reference spectra in the database. The area of the peaks from a metabolite is directly related to its abundance. As the abundance changes, the heights of peaks belonging to the same metabolite vary, as shown in the case of taurine in Figure 1. This enables relative quantification of metabolites across samples of

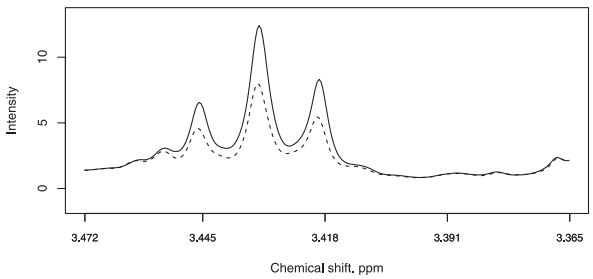

Fig. 1. A region from two ^{1}H NMR spectra from the spike-in experiment in Section 4. Solid line: spectral region from a sample containing the spiked metabolite taurine in the concentration of 3200 mols/l. Dashed line: spectral region from a sample containing taurine in the concentration of 1600 mols/l. The decrease in metabolite abundance results in a proportional decrease of all three peaks.

various types, e.g. across samples from patients with various status of a disease in biomarker discovery investigations.

However, inference of identity and abundance of metabolites from ^{1}H NMR spectra is fraught with difficulties. First, complex biological mixtures contain hundreds of metabolites, which vary in abundance substantially, and can produce highly overlapped peaks. Second, the number of candidate metabolites in a database typically exceeds the number of major sources of signals in the spectra, and one has to explore a combinatorially large space of candidate metabolites. Finally, due to numerous factors such as potential of pH and ionic interactions, peaks generated by a metabolite can deviate from their expected p.p.m. (Brelstaff *et al.*, 2009), introducing 'positional uncertainty'. The combination of these factors makes the deconvolution and the interpretation of individual metabolite-specific signals very challenging in practice. Automated inferential approaches are needed in order to account for the noise and ambiguity of spectral interpretations, and to derive objective, accurate and reproducible identifications and quantifications of metabolites.

This article develops a fully automated inferential approach, that we call *BQuant*, for identification and quantification of metabolites in regions of complex ^{1}H NMR spectra. It takes as input a set of NMR spectra, a range of p.p.m. of interest and a database of reference spectra for candidate compounds. It is based on a family of Bayesian hierarchical models for the spectra, where the models with the highest posterior probabilities are selected using a stochastic sampling technique. *BQuant* outputs a list of identified metabolites associated with their inclusion probability, and their measures of abundances in each spectrum on a continuous scale.

2 BACKGROUND

Once NMR spectra are acquired, their analysis involves baseline correction, peak linewidth adjustment, normalization, spectral alignment, identification and quantification of metabolites, and univariate and multivariate statistical analysis. This manuscript focuses on the identification and quantification of metabolites.

Methods for identification and quantification of metabolites can be classified into three broad categories: binning, curve-fitting without a database and curve-fitting with a database. For binning and curve-fitting without a database, feature quantification is performed first and identification of the features is conducted subsequently as a separate step. For curve-fitting with a database, identification and quantification are performed simultaneously.

Binning: this (Meyer *et al.*, 2008) processes one spectrum at a time, divides each spectrum into equally or variable-sized bins and integrates the intensities in each bin for quantification. The method is easy to implement, however it often lacks accuracy, in particular, in crowded regions of the spectra. Feature abundances can be inaccurately estimated by overlooking tail areas of peaks, or contaminated by signals from other sources.

The bins are manually annotated with metabolite identities using a database; however, this requires a great degree of expertise. Since multiple peaks of a metabolites are highly correlated, statistical total correlation spectroscopy (STOCSY) (Cloarec *et al.*, 2005) can be used to assist the identification (Alves *et al.*, 2009).

Curve fitting without a database: after phase and baseline correction, many NMR spectra can be viewed as a combination of multiple non-negative source signals from individual metabolites, mixed linearly with non-negative but unknown proportions. Curve fitting attempts to recover these signals directly from the entire collection of spectra and a database. The general task of curve fitting can be performed by a class of non-negative matrix factorization (NMF) methods. One such algorithm, Bayesian spectrum decomposition (Ochs *et al.*, 1999), sets specific forms of prior distributions for both source signals and mixing coefficients, and has been successfully applied to deconvolution of NMR spectra (Stoyanova *et al.*, 2004). Another NMF algorithm, alternating least squares (Eads *et al.*, 2004), iteratively estimates source signals and mixing coefficients until convergence. Finally, 'constrained total-line-shape fitting' (Laatikainen *et al.*, 1996; Soininen *et al.*, 2005) utilizes detected peak locations, and models the peaks by a linear combination of Lorentzian and Gaussian lineshape functions. Similar settings appear in Alsmeyer and Marquardt (2004) and Anderson *et al.* (2009).

As in the case of binning, identification of signals derived with curve-fitting methods occurs in a subsequent separate step. Unfortunately, the extracted curves are often not interpretable directly, and this undermines the practical utility of these methods.

Curve-fitting with database: interpretation of ^{1}H NMR spectra as linear combinations of reference profiles from a database (Wishart, 2008) allows us to simultaneously identify and quantify metabolites. Crockford *et al.* (2005) utilize recorded pure compound spectra as signal templates and describe a least-squares strategy for fitting the templates to spectra. Gipsona *et al.* (2006) utilize the database while imposing non-negativity constraints on the relative abundance of metabolites, and show a promising constrained fitting scheme called least-squares deconvolution. However, the approaches have been demonstrated in limited experimental situations.

Finally, the commercial Chenomx NMR suite provides a comprehensive database of metabolites, and their signal curves consisting of Lorentzian peaks, that can be used for manual deconvolution. Figure 2 provides an example of Chenomx analysis of an ^{1}H NMR spectrum from human plasma samples in Section 4. The raw spectrum (in black) is decomposed into multiple components and their summation is depicted by another black curve below the raw spectrum. The shaded area indicates four peaks, which are generated by a metabolite valine according to the database in Chenomx. Despite its manual nature, Chenomx suite is currently a popular tool for the interpretation of NMR metabolomic spectra.

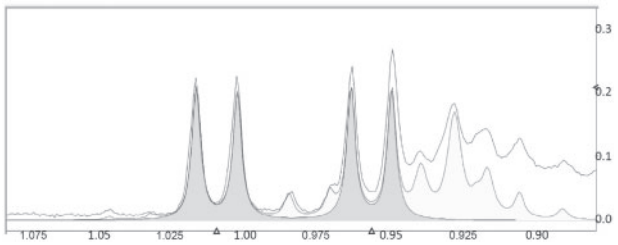

Fig. 2. Interpretation of an ^{1}H NMR spectrum from the the metabolomic investigation of plasma samples in Section 4 using Chenomx. *X*-axis: p.p.m.; *y*-axis: intensity; two black curves: raw spectrum and fitted spectrum; shaded area: four peaks generated by metabolite valine, according to a reference database. The difference between the observed spectrum and the fitted spectrum is primarily due to signals from large molecules such as lipids.

In summary, simple methods such as binning oversimplify the structure of the spectra and can produce inaccurate identification and quantification. Curve fitting without a database uses realistic models, but often fails to capture the true signals. Database-based approaches improve both the interpretability and the accuracy of spectral interpretation; however, they lack automation.

This article proposes an automated probabilistic approach for curve fitting using a database of candidate metabolites. Similar to the existing methods, we model the observed spectra as a combination of reference signals in a database, mixed linearly with unknown proportions. However, unlike the existing methods, we specify a Bayesian probability model which jointly represents the full set of replicate spectra and views metabolites in the database as candidate variables. We then convert the problem of metabolite identification into a task of variable selection, and develop a stochastic variable searching scheme that explores the model space and identifies highly likely metabolites. Results are delivered in a format similar to Chenomx.

3 METHODS

3.1 Preprocessing

We apply the following processing steps to raw ^{1}H NMR spectra: (i) a simple linear baseline is estimated and subtracted from each spectrum (Craig *et al.*, 2006); (ii) reference deconvolution is applied to adjust the linewidth of the peaks (Metz *et al.*, 2000); (iii) the spectra are normalized by the total sum normalization (Zhang *et al.*, 2009); (iv) the spectra are aligned by recursive segment-wise peak alignment in (Veselkov *et al.*, 2009). All other approaches for baseline correction, normalization, spectrum alignment and downstream statistical analysis can be used in conjunction with *BQuant*.

3.2 Database of candidate metabolites

For experimental datasets, we use an in-house database of experimental spectra from 290 pure compounds. The spectra were acquired using a 600 MHz spectrometer, NOESY-presaturation sequence with a mixing time of 100 ms, prescan delay of 1.00 s, 4 s acquisition time and sample temperature 25°C. Samples were prepared in H_2O (with 10% D_2O, at pH = 6.5). For each metabolite, the database specifies a reference ^{1}H NMR spectrum, the number of protons, the number of peaks derived from each proton and the information on locations, linewidths and relative heights of the peaks. Other databases, in particular the database available from Chenomx NMR suite, can be used directly with the proposed approach.

3.3 BQuant: estimation of peak shifts

Locations of peaks in the spectra can deviate from the expected p.p.m. in the database, even after the spectra are well aligned. This 'positional uncertainty' is experiment specific, and reference spectra need to be adjusted anew for each dataset. Patterns of 'positional uncertainty' are defined by the chemical structures of the metabolites, and follow specific rules. For example, the first two valine peaks in Figure 2 are generated by the same proton and form a doublet, and the last two valine peaks form another doublet. All peaks in a multiplet can shift to the left or to the right simultaneously by the same amount; however, the relative distances between multiplets vary.

We view shifts between the expected and the observed peak locations as unknown parameters, and estimate them from the spectra separately for each metabolite. We then employ a straightforward heuristic searching scheme for estimating peak location shifts, which amounts to solving a series of restricted least-squares optimization problems by quadratic programming. A more formal description of the approach is provided in the Supplementary Materials.

3.4 BQuant: reduction of the observed spectra

The raw experimental spectra contain a large number of data points, stored in files of up to 32 000 points per spectrum. To reduce the computational burden, we select a local region of interest. We exclude from consideration the region surrounding 4.7 p.p.m., where signals from water dominate all other signals from metabolites. We then replace the full-resolution spectra with a list of fixed peak locations, derived using a peak picking approach such as in Zhang *et al.* (2009), as well as with the list of local minima between peaks, defined as the positions of bin edges determined by the Adaptive Intelligent Binning algorithm (Meyer *et al.*, 2008).

3.5 BQuant: probability model for the spectra

Bayesian hierarchical model: suppose that we have observed n ^{1}H NMR spectra from n biological samples. We assume that all n samples are selected from the same homogeneous population, and do not contain systematic differences in abundances of metabolites, e.g. due to disease or treatment groups. When the study contains multiple groups, *BQuant* is applied separately to each group.

Suppose that each spectrum records intensities at peak locations or bin edges at J positions in total. Let vector $\boldsymbol{y_i}$ denote the intensities of spectrum i at these J positions and $\boldsymbol{y} = (\boldsymbol{y_1}', \ldots, \boldsymbol{y_i}', \ldots, \boldsymbol{y_n}')'$ denote the vector of length $J \times n$ for all the spectra intensities from n spectra combined. Also suppose that we have a database with reference spectra of L metabolites, adjusted for 'positional uncertainty'. Let vector $\mathbf{x}_l$ denote the intensities of the reference spectrum of metabolite l at the same J positions. Let $\boldsymbol{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_l, \ldots, \mathbf{x}_L)$ be the $J \times L$ design matrix formed by all the reference spectra combined.

We model the spectra as a linear combination of signals from multiple sources. First, large molecules such as lipids can produce broad NMR peaks, which may be viewed as parts of the baseline but overlooked by the preprocessing steps. We model such non-specific signals using a penalized cubic B-spline. We denote as K the number of basis functions and as $\boldsymbol{S}$ the $J \times K$ design matrix corresponding to the cubic B-spline basis (Eilers and Marx, 1996) and set the knots of the B-splines at locations of bin edges.

The remaining signals are generated by the metabolites in a database, and we distinguish signals of systematic and of stochastic nature. Each systematic source of signals represents the average relative strength of signal from a metabolite, as compared with the other metabolites in all samples. The stochastic source of signals reflects the biological variation of the metabolite abundance across samples and the measurement error. Therefore, the overall spectral signal $\boldsymbol{y_i}$ from a sample i can be represented as a linear combination of signals from these sources:

$$\boldsymbol{y_i} = \boldsymbol{S\phi} + \boldsymbol{X_\gamma \beta_\gamma} + \boldsymbol{X_\gamma \alpha_{i,\gamma}} + \boldsymbol{\epsilon_i} \quad (1)$$

$$\boldsymbol{\alpha_{i,\gamma}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_{\alpha;\gamma}})$$

$$\boldsymbol{\epsilon_i} = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{iJ}), \text{ and } \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$
$$i = 1, 2, \ldots, n.\ j = 1, 2, \ldots, J.$$

where $x \sim D$ indicates that the random variable x has the probability distribution D. The mixing parameters are the K-vector $\boldsymbol{\phi}$ of coefficients of the spline, the L-vector $\boldsymbol{\beta}$ of average relative abundances of the metabolites across all the samples and the L-vectors $\boldsymbol{\alpha_i}$ of deviations of abundances of the metabolites in sample i from the overall average levels. Parameter $\boldsymbol{\beta}$ is the same for all the samples, and $\boldsymbol{\alpha_i}$ are sample-specific random variables, and can be referred to, respectively, as fixed and random effects. In other words, the overall signal in a spectrum is represented by a linear mixed effects model. $\boldsymbol{\alpha_i}$ define a hierarchical structure of variation through their own probability distribution and covariance $\boldsymbol{\Sigma_\alpha}$. Since we simultaneously consider the relative strength of the signal between different metabolites in all samples, and the relative abundance of a metabolite across samples, the design matrix is the same for both fixed and random effects.

Only a subset of the metabolites in the database is present in the sample and can produce signals. Therefore, Equation (1) defines a family of models. Different instances of the family can be obtained by means of indicator variable $\boldsymbol{\gamma}$, which specifies the inclusion of individual metabolites. Specifically, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_l, \ldots, \gamma_L)$ defines the design submatrix $\boldsymbol{X_\gamma}$, the fixed and random effects $\boldsymbol{\beta_\gamma}$ and $\boldsymbol{\alpha_{i,\gamma}}$, and the variance-covariance of the random effects $\boldsymbol{\Sigma_{\alpha;\gamma}}$, which correspond to a particular subset of metabolites in the database. Therefore, in the Bayesian framework, identifying metabolites in the experiment is equivalent to searching the space of candidate models (i.e. the space of indicator variables $\boldsymbol{\gamma}$) and finding models which have high posterior probabilities.

Prior distributions of model parameters: due to the necessary adjustments for 'positional uncertainty', columns of the design matrix $\boldsymbol{X}$ tend to be correlated. Consequently, the posterior distribution of the candidate models can be relatively flat. At the same time, we are not interested in all the candidate modes, but only in a subset of sparse models which explain the spectra with a minimal number of abundant metabolites. We reflect our preference for sparsity by a second indicator variable $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_l, \ldots, \delta_L)$, which determines the practically significant presence of the metabolites. The indicator controls the prior distributions of β and α_i, helps obtain a more peaked posterior distribution of $\boldsymbol{\gamma}$ and enables inference. The setting is motivated by the stochastic search variable selection for linear regression proposed by George and McCulloch (1993) and extended by O'Hara and Sillanpaa (2009). Additional discussion on this setting is provided in Supplementary Materials. The priors for $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are

$$\gamma_l \overset{\text{iid}}{\sim} \mathcal{B}(p_{1,l}),\ \delta_l \overset{\text{iid}}{\sim} \mathcal{B}(p_{2,l}),\ l = 1, \ldots, L.$$

where $\mathcal{B}(\cdot)$ denotes a Bernoulli distribution, and both hyperparameters $p_{1,l}$ and $p_{2,l}$ are predefined constants. $p_{1,l}$ reflects our prior knowledge on the inclusion probability of metabolite l, and larger values of $p_{1,l}$ yield more selected metabolites. $p_{2,l}$ reflects our prior view of the practical significance of the average abundance of metabolite l. For reasons of computational tractability, we specify conjugate prior distributions of $\boldsymbol{\phi}$, components of the vector $\boldsymbol{\beta}$, submatrices of $\boldsymbol{\Sigma_\alpha}$ and σ^2 as follows:

$$p(\boldsymbol{\phi}) \propto \mathcal{N}(\mathbf{0}, \tau_\phi \mathrm{I_K}) \cdot \exp(-\frac{1}{2}\lambda \boldsymbol{\phi}' \boldsymbol{P} \boldsymbol{\phi}) \quad (2)$$

$$p(\beta_l | \gamma, \delta) \propto \mathcal{N}(0, \nu_1)\, I(\beta_l \geq 0)\, I(\delta_l = 1)\, I(\gamma_l = 1) + \mathcal{N}(0, \nu_0)\, I(\beta_l \geq 0)\, I(\delta_l = 0)\, I(\gamma_l = 1),$$
$$l = 1, \ldots, L \quad (3)$$

$$\boldsymbol{\Sigma_{\alpha;\gamma,\delta}} | \gamma, \delta \sim \mathcal{IW}(\mu_{0;\gamma,\delta}, \boldsymbol{\Sigma_{0;\gamma,\delta}}) \quad (4)$$

$$\sigma^2 \sim \mathcal{IG}(a_\sigma, b_\sigma) \quad (5)$$

Here $p(x) \propto f(x)$ indicates that the density function for the probability distribution of x is proportional to $f(x)$, i.e. is specified up to a constant. $\mathrm{I_K}$ is the $K \times K$ identity matrix, $I(\cdot)$ is indicator function. $\mathcal{IW}(\mu_{0;\gamma,\delta}, \boldsymbol{\Sigma_{0;\gamma,\delta}})$ is the Inverse Wishart distribution with degree of freedom $\mu_{0;\gamma,\delta}$ and scale matrix $\boldsymbol{\Sigma_{0;\gamma,\delta}}$ and $\mathcal{IG}(a_\sigma, b_\sigma)$ is the Inverse Gamma distribution with shape a_σ and scale b_σ.

The prior distribution of $\boldsymbol{\phi}$ in Equation (2) follows the proposal by Wood (2006), which views the baseline as an extension of the original penalized B-spline (Eilers and Marx, 1996). The exact form of matrix $\boldsymbol{P}$ is given in Supplementary Materials. The hyperparameter λ in this distribution controls the smoothness of the baseline. We select λ via an Empirical Bayes estimation step, where we fit the model in Equation (1) to the constructed spectrum with median intensities as the response, and estimate λ automatically by generalized cross-validation (GCV) (Golub *et al.*, 1979).

The prior for each element β_l is conditional on metabolite inclusion. It is specified as a Truncated Normal distribution, reflecting the positivity constraints on the average abundance of the metabolites. The priors for $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are a direct extension of the setting proposed by Thompson and Rosen (2008) for variable selection in linear mixed models. All the components of $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are assumed to be independent *a priori*. Hyperparameters ν_0 and ν_1 are predefined constants, as in George and McCulloch (1993).

Equation (4) specifies separate prior distributions for the variance-covariance matrices of the subset of metabolites in the database $\boldsymbol{\Sigma_{\alpha;\gamma,\delta}}$, conditional on metabolite inclusion. The prior for $\boldsymbol{\Sigma_{\alpha;\gamma,\delta}}$ is made non-informative as in Kass and Natarajan (2006), by assigning $\mu_{0;\gamma,\delta} = \Sigma_{l=1}^{L}(\gamma_l \cdot \delta_l)$, and the sample variance-covariance matrix to $\boldsymbol{\Sigma_{0;\gamma,\delta}}$.

Finally, the prior of σ^2 in Equation (5) is made non-informative by assigning a_σ and b_σ to small values such as 0.001. Choice of all the parameters is discussed in more details in Supplementary Materials.

Metabolite identification as a variable selection task: the problem of metabolite identification is similar to Bayesian variable selection in linear regression (Chipman *et al.*, 2001; George and McCulloch, 1997), and is complicated by the inclusion of subject-specific random effects. It is also similar to variable selection developed for Bayesian linear mixed models and generalized linear models (Cai and Dunson, 2006; Chen and Dunson, 2003; Kinney and Dunson, 2007), where we automatically identify subsets of predictors having non-zero coefficients of fixed effects or non-zero variances of random effects. However, such approaches are not directly applicable to the model in (1)–(5), as they treat fixed and random effects separately, and do not allow a simultaneous selection of variables that are both fixed and random effects.

An alternative variable selection approach has been proposed for automatic selection of knot locations for B-spline, where a linear mixed model restricts fixed and random effects to be the same set of variables representing knot locations (Thompson and Rosen, 2008). Although this approach addresses a similar variable selection task, the model in (1)–(5) is more comprehensive. It contains extra fixed effects for baseline, non-negativity constraints on the parameter space and additional indicator variables for practical significance controlling both fixed and random effects. Therefore, we propose a Gibbs sampling scheme to explore the high-dimensional variable space, to obtain joint posterior probabilities for sets of parameters, as well as marginal posterior probabilities for the statistical and practical significance of individual metabolites.

3.6 BQuant: Bayesian inference

Gibbs sampling scheme: denote $\boldsymbol{\alpha} = (\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, \ldots, \boldsymbol{\alpha_n})$ the matrix containing all the random effects for all the samples, and $\boldsymbol{\alpha_\gamma} = (\boldsymbol{\alpha_{1,\gamma}}, \boldsymbol{\alpha_{2,\gamma}}, \ldots, \boldsymbol{\alpha_{n,\gamma}})$ the submatrix of $\boldsymbol{\alpha}$ where rows correspond to non-zero elements in $\boldsymbol{\gamma}$. $\boldsymbol{\alpha_{\gamma,\delta}}$ is a submatrix of $\boldsymbol{\alpha}$ where rows with correspond to non-zero elements in both $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$. The joint posterior distribution for $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\phi}, \boldsymbol{\beta_\gamma}, \boldsymbol{\alpha_\gamma}, \boldsymbol{\Sigma_{\alpha;\gamma}}, \sigma^2)$ is:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto$$
$$p(\boldsymbol{y}|\boldsymbol{\theta})\, p(\boldsymbol{\gamma}, \boldsymbol{\delta})\, p(\boldsymbol{\phi})\, p(\boldsymbol{\beta_\gamma}|\boldsymbol{\gamma}, \boldsymbol{\delta})\, p(\boldsymbol{\alpha_\gamma}|\boldsymbol{\Sigma_{\alpha;\gamma}})\, p(\boldsymbol{\Sigma_{\alpha;\gamma}}|\boldsymbol{\gamma}, \boldsymbol{\delta})\, p(\sigma^2)$$

The Gibbs sampling scheme iteratively samples from the following conditional distributions.

(1) Sample fixed effects $(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\phi})$ from their conditional distribution:

$$p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\phi} | \gamma, \boldsymbol{\Sigma}_{\boldsymbol{\alpha};\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{y}) \propto \mathcal{MVN}\left((\boldsymbol{\mu}_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}, \boldsymbol{\mu}_{\boldsymbol{\phi}}), \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\phi}}\right) \cdot I_{q(\gamma)}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \geq \mathbf{1}_{q(\gamma)} 0\right) \quad (6)$$

where $\mathcal{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ and $I_{q(\gamma)}(\boldsymbol{V_1} \geq \boldsymbol{V_2})$ denotes a component-wise indicator function. The expressions for $\boldsymbol{\mu}_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}, \boldsymbol{\mu}_{\boldsymbol{\phi}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\phi}}$ are given in Supplementary Materials. The joint posterior distribution of $(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\phi})$ is a Truncated Multivariate Normal distribution, primarily due the non-negativity constraint on $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. Sampling from this distribution requires a separate Gibbs sampling scheme, such as the one in Kotecha and Djuric (1999). We follow this by the sampling of $\boldsymbol{\phi}$ drawn from a Multivariate Normal distribution conditional on $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$.

(2) Sample random effects $\boldsymbol{\alpha}_{\boldsymbol{\gamma}}$ by drawing instances of all $\boldsymbol{\alpha}_{i,\boldsymbol{\gamma}}$ from their conditional distributions $p(\boldsymbol{\alpha}_{i,\boldsymbol{\gamma}} | \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{\Sigma}_{\boldsymbol{\alpha};\boldsymbol{\gamma}}, \boldsymbol{y}) = \mathcal{MVN}(\boldsymbol{\mu}_{i,\boldsymbol{\alpha};\boldsymbol{\gamma}}, \boldsymbol{V}_{i,\boldsymbol{\alpha};\boldsymbol{\gamma}})$ where $\boldsymbol{\mu}_{i,\boldsymbol{\alpha};\boldsymbol{\gamma}}$ and $\boldsymbol{V}_{i,\boldsymbol{\alpha};\boldsymbol{\gamma}}$ are given in Supplementary Materials.

(3) Sample the variance of measurement error σ^2 from $p(\sigma^2 | \boldsymbol{\alpha}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\phi}, \boldsymbol{y}) = \mathcal{IG}(c_\sigma, d_\sigma)$ where c_σ and d_σ are given in Supplementary Materials.

(4) Sample the variance-covariance matrix for random effects $\boldsymbol{\Sigma}_{\boldsymbol{\alpha};\boldsymbol{\gamma},\boldsymbol{\delta}}$

$$p(\boldsymbol{\Sigma}_{\boldsymbol{\alpha};\boldsymbol{\gamma},\boldsymbol{\delta}} | \boldsymbol{\alpha}, \boldsymbol{y}) = \mathcal{IW}(\mu_{0;\gamma,\delta} + n, \boldsymbol{\Sigma}_{\mathbf{0};\boldsymbol{\gamma},\boldsymbol{\delta}} + \boldsymbol{\alpha}_{\boldsymbol{\gamma},\boldsymbol{\delta}}{}' \boldsymbol{\alpha}_{\boldsymbol{\gamma},\boldsymbol{\delta}})$$

where $\boldsymbol{\alpha}_{\boldsymbol{\gamma},\boldsymbol{\delta}}$ is the submatrix of $\boldsymbol{\alpha}$ where only the rows corresponding to both $\gamma_l = 1$ and $\delta_l = 1$ are retained.

(5) Sample $\boldsymbol{\gamma}$ from $\mathcal{B}(p(\gamma_l = 1 | \boldsymbol{\gamma}_{(-l)}, \boldsymbol{\delta}, \sigma^2, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}, \boldsymbol{y}))$ where $\boldsymbol{\gamma}_{(-l)}$ is equal to $\boldsymbol{\gamma}$ excluding the l-th component. $p(\gamma_l = 1 | \boldsymbol{\gamma}_{(-l)}, \boldsymbol{\delta}, \sigma^2, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}, \boldsymbol{y})$ is given in Supplementary Materials. Only one component of $\boldsymbol{\gamma}$ is updated at this step for the current iteration and a different component will be updated in the next iteration until every component of $\boldsymbol{\gamma}$ is updated once. Then the entire vector of $\boldsymbol{\gamma}$ will be sampled component-wise the second time.

(6) Sample $\boldsymbol{\delta}$, from $\mathcal{B}(p(\delta_l = 1 | \gamma_l = 1, \boldsymbol{\gamma}_{(-l)}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\Sigma}_{\boldsymbol{\alpha};\boldsymbol{\gamma}}, \boldsymbol{y}))$ The details of this distribution are given in Supplementary Materials.

Posterior inference: Gibbs sampling is continued $N_b + N_s$ iterations with a burn-in period of N_b. We focus on the posterior distributions of the parameters which are directly related to the identification and quantification of metabolites in (1). These parameters include $\boldsymbol{\gamma}$ for indicating the subset of the metabolites in the database that are present with high probability, β_l for average abundance of metabolite l and $\mu_{l,i} = \beta_l + \alpha_{l,i}$ for representing the abundance of metabolite l in sample i. The parameters can be estimated using two approaches. The first involves model averaging (Madigan and Raftery, 1994), where the parameters are estimated by their posterior means and approximated by averaging over all draws of Gibbs sampling after the initial burn-in period:

$$\hat{\boldsymbol{\gamma}} = \frac{1}{N_s} \sum_{s=N_b+1}^{N_b+N_s} \boldsymbol{\gamma}^{(s)}; \qquad \hat{\boldsymbol{\beta}} = \frac{1}{N_s} \sum_{s=N_b+1}^{N_b+N_s} \boldsymbol{\beta}^{(s)}$$

$$\hat{\mu}_{l,i} = \frac{1}{N_s} \sum_{s=N_b+1}^{N_b+N_s} \left(\beta_l^{(s)} + \alpha_{l,i}^{(s)}\right) \quad (7)$$

Alternatively, we can estimate $\boldsymbol{\gamma}$ by the vector with largest posterior probability among all vectors visited by the Gibbs sampler (Sha *et al.*, 2006)

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}^{(s)}}{\operatorname{argmax}} \hat{p}(\boldsymbol{\gamma}^{(s)} | \boldsymbol{Y}) \quad (8)$$

where $\hat{p}(\boldsymbol{\gamma}^{(s)} | \boldsymbol{Y})$ is the relative frequency of $\boldsymbol{\gamma}^{(s)}$. Equation (7) will be primarily used for assessing the performance of *BQuant* in both identification and quantification. Models with high probability identified by Equation (8) will also be reported as a summary of the results.

4 DATASETS

Acquisition of NMR spectra: the experimental ^{1}H NMR spectra were obtained using the NOESYPR pulse sequence at 500 MHz on a Bruker Avance NMR spectrometer and utilizing a z-gradient HCN triple resonance cryoprobe. Extensive details are provided in Supplementary Materials.

Simulated spectra: we simulate $n = 20$ spectra with $J = 100$ intensities per spectrum. The simulation follows the model in Equation (1); however, additional noise signals are introduced to test for robustness. The simulated spectra are constructed from four out of a total of eight candidate metabolites.

Spike-in experiment: a urine sample was collected from a healthy male individual, according to a protocol approved by the Institutional Review Board at Purdue University. Five metabolites (taurine, hippuric acid, nicotinic acid, malic acid and oxoglutaric acid) were added to the urine sample in varying concentrations, according to a latin square design. These metabolites were chosen as they are potentially observable in urine with ^{1}H NMR; however, no noticeable abundance of these metabolites was detected in this specific urine sample. Therefore, the sample will have little interference on total abundances of these metabolites in the spiked mixtures. Three replicate ^{1}H NMR spectra were acquired from each mixture. For each of the five spiked metabolites, we manually select a local region containing pronounced peaks of the metabolite. The regions are listed in Table 3, and an example is shown in Figure 1.

Metabolomic investigation of plasma samples: we analyze plasma samples from 46 healthy individuals. The Institutional Review Board at Indiana University approved the study protocol, and written informed consent was obtained from all subjects before enrollment. Since the true identities and abundances of metabolites in these samples are unknown, we compare the ability of various methods to identify and quantify metabolites against a benchmark of manual identifications by NMR experts. Moreover, independent laboratory quantification of one metabolite in these samples (creatinine) was obtained as part of routine clinical care, and are available for all the 46 samples. Thus, we evaluate the performance of statistical methods in terms of agreement between the enzymatic (i.e. laboratory) and NMR-based quantifications of creatinine.

We select two regions to demonstrate the application of *BQuant*, which include 0.8852–1.04 p.p.m. and 3.92–4.3 p.p.m. The first region contains a moderate degree of peak overlap and a non-zero baseline. Peak deconvolution is, therefore, particularly important for identifying these metabolites. The second region contains the best-resolved creatinine peak. Although the shape of the peak is well preserved, its signal is convoluted with other signals, in particular with a broad lipid peak. These two regions represent examples of typical real-life challenges in spectral interpretation.

5 RESULTS

We compared the performance of *BQuant* to representative instances of three classes of fully automated methods in Section 2. Binning is represented by the adaptive intelligent binning (AI binning) (Meyer *et al.*, 2008). An instance of curve fitting without a database is the multiplicative update algorithm implemented in the R package 'Non-negative matrix factorization (`nmf`)' (Liu, 2009), which we abbreviate as MU-NMF. Both AI binning and MU-NMF cannot automatically identify the metabolites, and cannot be compared to *BQuant* in terms of identification. Therefore, we attempted to identify the bins (or the components of the deconvoluted mixtures) manually whenever possible, to compare the performance of these methods in terms of quantification.

Curve fitting with a database is represented by least-squares deconvolution (LD) (Gipsona *et al.*, 2006), and enables a simultaneous identification and quantification of metabolites. Since the original implementation of LD does not account for a baseline, this can negatively affect the performance of the method. Hence

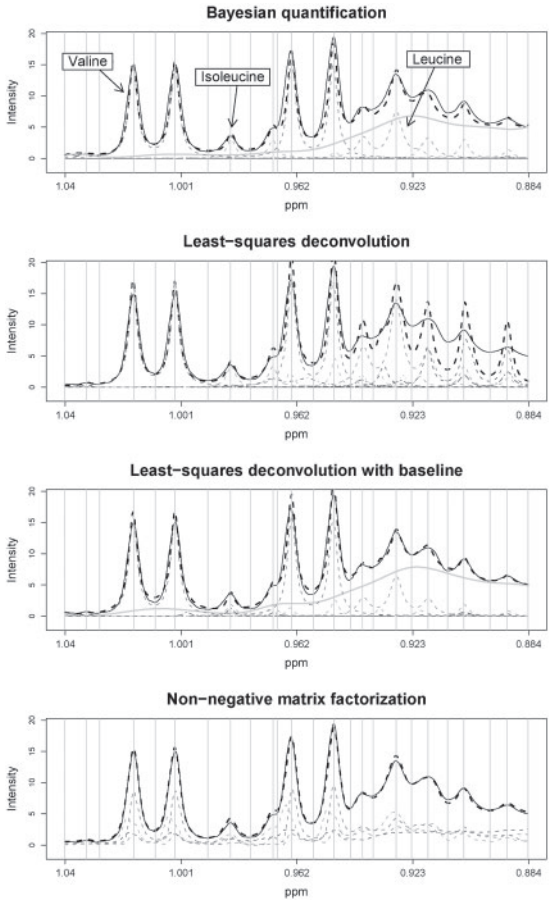

Fig. 3. Results of deconvolution of the spectrum in Figure 2. Black solid line: original spectrum; black dashed line: fitted spectrum; flat grey solid line: fitted baseline; dashed lines in different grey level: each curve corresponds to one component (one metabolite). The peaks corresponding to the three major metabolites are annotated in the first plot.

we also implemented a modified version of LD which, similarly to *BQuant*, incorporates a sample-specific baseline using penalized B-spline. This is accomplished by adding a term including the same design matrix $\boldsymbol{S}$ as specified in model (1), and a sample-specific coefficient ϕ_i, to the original formulation of LD. We refer to this modified version as LD with baseline (LDBL). We applied both LD and LDBL to the individual spectra in full resolution. Since the methods are only applicable to the individual spectra, we compared them to *BQuant* according to the list of most frequently identified metabolites.

To reduce the computational complexity of *BQuant*, we used LDBL for reduction of the size of the database. We chose $L_{\max} = 35$, since local regions typically do not contain more than 35 significantly abundant metabolites.

5.1 Evaluation on the simulated spectra

Identification accuracy: our goal was to identify the four true metabolites out of the total of eight candidate metabolites. For LD and LDBL, the model was fit to each spectrum individually and we inspected the top four metabolites with the highest estimated abundance. For *BQuant*, according to Equation (8), the model with the highest posterior probability was the true model with four out of eight metabolites selected. Table 1 compares the accuracy of automated metabolite identification by LD, LDBL and BQuant and shows that identification by *BQuant* outperforms these methods in terms of accuracy of identification.

Table 1. Identification accuracy on the simulated dataset

Number of samples	LD	LDBL	BQuant
All correct identifications/total	0/20	14/20	20/20

Each column shows the number of samples (out of 20 total), in which the methods correctly identify all 4 simulated metabolites.

Table 2. Quantification accuracy on the simulated dataset

R^2	AI Binning	LD	LDBL	BQuant
M_1	0.915	0.609	0.961	0.975
M_2	0.971	0.993	0.993	0.987
M_3	0.995	0.987	0.983	0.998
M_4	0.977	0.984	0.997	0.996

Each row corresponds to one metabolite. The four columns list the R^2 values for a simple linear regression with no intercept, where the true abundance is used as the predictor, and the quantified abundance as the response variable. A higher value indicates better performance.

Quantification accuracy: to evaluate the quantification accuracy of AI binning, we manually selected bins dominated by signals from the true simulated standard spectra. Supplementary Materials show an example of MU-NMF deconvolution of one simulated spectrum. It illustrates that the identified components are not immediately interpretable in terms of the metabolites in the database, and cannot be used for quantification.

Successful quantifications should be linearly associated to the theoretical abundance of the metabolites, with a 45-degree slope and no intercept. The R^2 of this linear model, fitted separately for each method of metabolite quantification, are summarized in Table 2. Overall, *BQuant* showed the best performance. A closer case-by-case inspection revealed that metabolites with fewer peaks and lower relative abundance were more difficult to identify and quantify by all methods.

5.2 Evaluation on the spike-in dataset

Identification accuracy: LD, LDBL and *BQuant* could all identify the five spiked metabolites from the spectra.

Quantification accuracy: as in the case of simulated dataset, we manually identified the bins resulting from the AI binning; however, no manual identification could be obtained for the output of MU-NMF. After fitting a linear regression model with no intercept, the true abundance as the predictor and the quantified abundance as the response for each method, the R^2s are summarized in Table 3. Overall, *BQuant* outperformed the other three methods. The suboptimal performance of AI Binning was primarily due to the contributions of tail areas from neighboring peaks to the abundance of the bins. Since LD decomposes the spectrum individually, and does not synthesize the information across samples, it underperformed *BQuant* in regions with heavy overlap.

Table 3. Quantification accuracy on the spike-in experiment

R^2	Region	AI Binning	LD	LDBL	BQuant
Taurine	3.37–3.47	0.861	0.994	0.986	0.987
Hippuric acid	7.5–7.88	0.728	0.769	0.778	0.794
Nicotinic acid	8.2–8.404	0.975	0.991	0.999	0.999
Malic acid	2.34–2.43	0.754	0.829	0.919	0.983
Oxoglutaric acid	2.42–3.02	0.824	0.638	0.925	0.970

Each row corresponds to one spike-in metabolite. The first column shows the selected regions of the spectra in p.p.m. The other four columns list the R^2 values for a simple linear regression with no intercept, where the true abundance is used as the predictor, and the quantified abundance as the response variable. A higher value indicates better performance.

5.3 Application to the investigation of plasma samples

Identification accuracy: the goal of the investigation is to accurately identify and quantify metabolites in regions of ^{1}H NMR spectra from plasma samples of 46 subjects. The first region of interest in this dataset is 0.8852–1.04 p.p.m. Manual identification of metabolites, aided by Chenomx, reveals three major metabolites (leucine, isoleucine and valine), and a local baseline created by the lipid signal. Figure 3 illustrates the deconvolution of one of the spectra in the region by MU-NMF, LD, LDBL and *BQuant*. As can be seen in the figure, the components extracted by MU-NMF were not immediately interpretable. They also conflicted with manual interpretations of the spectra. The results of LD are severely hampered by the presence of the baseline, which is not properly accounted for by the model. Although leucine and valine were identified as the two most abundant metabolites across all the samples, isoleucine was not among the top three most abundant metabolites in any of the 46 cases. The results of LDBL have a higher agreement with the manual identification; however, isoleucine remained excluded from the list of three most abundant metabolites in 15 out of 46 samples. This is due to a small peak near 0.98 p.p.m., which was falsely decomposed as a sum of peaks from two separate metabolites (butanone and isoleucine), instead of from isoleucine alone. As shown in Table 4, *BQuant* contains leucine, isoleucine and valine as components in the most probable *a posteriori* model, and assigns them the largest estimates of abundance. It is worth mentioning that manual identification of metabolites such as 2-aminobuyrate is challenging, as they typically have a weak signal in a region that is occupied by other metabolites.

Similarly, imprecise results were obtained in the second region of 3.92–4.3 p.p.m. with MU-NMF, LD and LDBL. Table 4 shows that *BQuant* automatically identified the metabolites, which were manually confirmed by NMR experts.

Quantification accuracy: as in the previous cases, we manually identified the bins resulting from the AI binning, but could not identify the signals from the output of MU-NMF, to evaluate their accuracy of quantification. Table 5 shows the coefficients of multiple determination of the abundance of creatinine obtained by enzymatic methodology , and derived by AI binning, LD, LDBL and *BQuant* from the NMR spectra. Since creatinine has two peaks, two bins are listed separately for AI binning, and the first bin contains signals from additional overlapped peaks. As can be seen, *BQuant* yields the best quantification result among all the methods.

Table 4. Identification and quantification of metabolites by *BQuant* in the metabolomic investigation of plasma samples

Region	Metabolite	Mean	95% CI		γ	δ	Confirm
0.8852–1.04	Valine	49.61	46.85	52.57	1.00	0.99	Yes
0.8852–1.04	Leucine	20.05	16.80	23.28	1.00	1.00	Yes
0.8852–1.04	Isoleucine	9.39	5.46	11.07	1.00	0.99	Yes
0.8852–1.04	2-ami*	3.52	2.27	4.42	1.00	0.92	No
0.8852–1.04	Cholate	0.96	0.00	3.69	0.57	0.21	No
3.92–4.3	Lactate	15.27	0.00	19.81	0.90	0.90	Yes
3.92–4.3	Creatinine	7.98	0.00	10.81	0.89	0.88	Yes
3.92–4.3	Serine	6.57	0.00	9.13	0.85	0.85	Yes
3.92–4.3	Cysteine	2.50	0.00	6.61	0.54	0.50	No
3.92–4.3	Threonine	2.27	0.00	3.71	0.77	0.65	Yes
3.92–4.3	3-hyd*	1.94	0.00	3.84	0.73	0.52	Yes
3.92–4.3	Proline	1.73	0.00	4.88	0.72	0.42	Yes
3.92–4.3	Cholate	1.51	0.00	5.66	0.42	0.34	No
3.92–4.3	2′-deo*	0.65	0.00	3.92	0.66	0.13	No
3.92–4.3	AMP	0.54	0.00	2.59	0.36	0.12	No

The table shows all the metabolites with the highest posterior probability, along with their posterior mean and 95% credible interval. $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are frequencies of $\gamma_l = 1$ and $\delta_l = 1$ after the burn-in period. In the second column, 2-ami* = 2-aminobutyrate; 3-hyd* = 3-hydroxybutyrate; 2′-deo* = 2′-deoxyguanosine. The last column indicates whether the metabolite identification is manually confirmed by NMR experts.

Table 5. Quantification accuracy of creatinine in the investigation of plasma samples, in terms of coefficient of multiple determination R^2 between enzymatic laboratory measurements and by NMR-based quantifications

Creatinine	AI Binning$_1$	AI Binning$_2$	LD	LDBL	BQuant
R^2	0.209	0.547	0.477	0.475	0.718

Higher values indicate better performance. AI Binning$_1$ corresponds to the integral between bin edges 3.0012–3.0196 p.p.m.; AI Binning$_2$ corresponds to the integral between bin edges 4.0109 –4.0329 p.p.m.

6 CONCLUSION

Database-based identification and quantification of metabolites from ^{1}H NMR spectra attracts growing attention. However, there is currently no fully automatic way of assisting this labor-intensive procedure. We developed a probabilistic approach *BQuant*, which combines linear mixed modeling and Bayesian model selection, and automatically identifies and quantifies metabolites in local regions of spectra. At the same time, the model used in *BQuant* is more complex than the standard linear mixed models, as it accounts for the specialized characteristics of NMR spectra. To the best of our knowledge, *BQuant* is the first application of linear mixed modeling to the database-based ^{1}H NMR spectra deconvolution.

BQuant outperformed the existing automated approaches in terms of accuracy of both identification and quantification of metabolites. The improved performance is achieved for two reasons. First, the problems of identification and quantification of metabolites are interrelated, in particular in regions with heavy peak overlap. Unlike many other methods, which solve these problems sequentially, *BQuant* identifies and quantifies the metabolites simultaneously and iteratively, by means of Gibbs sampling. Therefore, accurate identification of peaks helps improve the accuracy of quantification, and well-quantified metabolites help identify the sources of the

additional peaks. The second reason for the improved performance of *BQuant* is that, unlike many other approaches, it simultaneously models the entire collection of spectra. This allows us to incorporate the maximal amount of information into parameter estimation and inference. When the set of spectra is highly heterogeneous, BQuant could potentially produce less accurate results. However, such a case will be challenging for any automatic approach. Finally, the implementation of *BQuant* is practical and useful, and can be easily integrated with other automated data analysis pipelines.

Funding: National Institutes of Health (NIH) (grants R01GM085291, 5K23RR019540 and UL1RR025761); Indiana University Signature Center Initiative; Indiana 21st Century Research & Technology Fund; Bisland Dissertation Fellowship from Purdue University (to C.Z.).

Conflict of Interest: none declared.

REFERENCES

Alsmeyer,F. and Marquardt,W. (2004) Automatic generation of peak-shaped models. *Appl. Spectrosc.*, **58**, 986–994.

Alves,A. *et al.* (2009) Analytic properties of statistical total correlation spectroscopy based information recovery in 1H NMR metabolic data sets. *Anal. Chem.*, **81**, 2075–2084.

Anderson,P. *et al.* (2009) Characterization of 1H NMR spectroscopic data and the generation of synthetic validation sets. *Bioinformatics*, **25**, 2992–3000.

Brelstaff,G. *et al.* (2009) Bag of peaks: interpretation of NMR spectrometry. *Bioinformatics*, **25**, 258–264.

Cai,B. and Dunson,D. (2006) Bayesian covariance selection in generalized linear mixed models. *Biometrics*, **62**, 446–457.

Chen,Z. and Dunson,D. (2003) Random effects selection in linear mixed models. *Biometrics*, **59**, 762–769.

Chen,H. *et al.* (2006) Combining desorption electrospray ionization mass spectrometry and nuclear magnetic resonance for differential metabolomics without sample preparation. *Rapid Commun. Mass Spectrom.*, **20**, 1577–1584.

Chipman,H. *et al.* (2001) The practical implementation of Bayesian model selection. *IMS Lect Notes Monogr. Ser.*, **38**, 67–131.

Cloarec,O. *et al.* (2005) Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Anal. Chem.*, **77**, 1282–1289.

Craig,A. *et al.* (2006) Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Anal. Chem.*, **78**, 2262–2267.

Crockford,D. *et al.* (2005) Curve-fitting method for direct quantitation of compounds in complex biological mixtures using 1H NMR: application in metabonomic toxicology studies. *Anal. Chem.*, **77**, 4556–4562.

Eads,C. *et al.* (2004) Molecular factor analysis applied to collections of NMR spectra. *Anal. Chem.*, **76**, 1982–1990.

Eilers,P. and Marx,B. (1996) Flexible smoothing with b-spline and penalties. *Stat. Sci.*, **11**, 89–102.

Ernst,R. *et al.* (1990) *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*. Oxford University Press, USA.

George,E. and McCulloch,R. (1993) Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.*, **88**, 881–889.

George,E. and McCulloch,R. (1997) Approaches for Bayesian variable selection. *Stat. Sin.*, **7**, 339–373.

Gipsona,G.T. *et al.* (2006) Weighted least-squares deconvolution method for discovery of group differences between complex biofluid 1H NMR spectra. *J. Magn. Reson.*, **183**, 269–277.

Golub,G. *et al.* (1979) Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215–223.

Gowda,G. *et al.* (2008) Metabolomics-based methods for early disease diagnostics: a review. *Exp. Rev. Mol. Diagn.*, **8**, 617–133.

Kass,R. and Natarajan,R. (2006) A default conjugate prior for variance components in generalized linear mixed models (comment on article by browne and draper). *Bayesian Anal.*, **1**, 535–542.

Kinney,S. and Dunson,D. (2007) Fixed and random effects selection in linear and logistic models. *Biometrics*, **63**, 690–698.

Kotecha,J. and Djuric,P. (1999) Gibbs sampling approach for generation of truncated multivariate gaussian random variables. *Proc. Acoust. Speech Signal Process.*, **3**, 1757–1760.

Laatikainen,R. *et al.* (1996) A computational strategy for the deconvolution of NMR spectra with multiplet structures and constraints: analysis of overlapping $^{1}3c$-^{2}h multiplets of ^{13}c enriched metabolites from cell suspensions incubated in deuterated media. *Magn. Reson. Med.*, **36**, 359–365.

Liu,S. (2009) *NMFN: Non-Negative Matrix Factorization*. R package version 1.0. 2009-11-24. Available at http://CRAN.R-project.org/package=NMFN

Madigan,D. and Raftery,A. (1994) Model selection and accounting for model uncertainty in graphical models using occam's window. *J. Am. Stat. Assoc.*, **89**, 1535–1546.

Metz,K. *et al.* (2000) Reference deconvolution: a simple and effective method for resolution enhancement in nuclear magnetic resonance spectroscopy. *Concepts Magn. Reson.*, **12**, 21–42.

Meyer,T.D. *et al.* (2008) NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Anal. Chem.*, **80**, 3783–3790.

Nicholson,J. *et al.* (1999) 'metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, **29**, 1181–1189.

Ochs,M. *et al.* (1999) A new method for spectral decomposition using a bilinear Bayesian approach. *J. Magn. Reson.*, **137**, 161–176.

O'Hara,R. and Sillanpaa,M. (2009) A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal.*, **4**, 85–118.

Sha,N. *et al.* (2006) Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics*, **22**, 2262–2268.

Silverstein,R. *et al.* (2005) *Spectrometric Identification of Organic Compounds*, 7th edn. Wiley, USA.

Soininen,P. *et al.* (2005) Strategies for organic impurity quantification by 1H NMR spectroscopy: constrained total-line-shape fitting. *Anal. Chim. Acta*, **542**, 178–185.

Stoyanova,R. *et al.* (2004) Sample classification based on Bayesian spectral decomposition of metabonomic NMR data sets. *Anal. Chem.*, **76**, 3666–3674.

Thompson,W. and Rosen,O. (2008) A Bayesian model for sparse functional data. *Biometrics*, **64**, 54–63.

Ulrich,E. *et al.* (2008). Biomagresbank. *Nucleic Acids Res.*, **36**, D402–D408.

Veselkov,K. *et al.* (2009) Recursive segment-wise peak alignment of biological 1H NMR spectra for improved metabolic biomarker recovery. *Anal. Chem.*, **81**, 56–66.

Weljie,A. *et al.* (2006) Targeted profiling: quantitative analysis of 1H-NMR metabolomics data. *Anal. Chem.*, **78**, 4430–4442.

Wishart,D. (2008) Quantitative metabolomics using NMR. *Xenobiotica*, **29**, 1181–1189.

Wishart,D. *et al.* (2007) HMDB: the human metabolome database. *Nucleic Acids Res.*, **35**, D521–D526.

Wood,S. (2006) *Generalized Additive Models: an Introduction with R*. Chapman and Hall, USA.

Zhang,S. *et al.* (2009) Interdependence of signal processing and analysis of urine 1H NMR spectra for metabolic profiling. *Anal. Chem.*, **81**, 6080–6088.