Biostatistics (2022) **00**, 00, *pp.* 1–16
https://doi.org/10.1093/biostatistics/kxac015

Shifting-corrected regularized regression for ^1H NMR metabolomics identification and quantification

THAO VU

Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Fitzsimons Building, 13001 East 17th Place, Aurora, CO 80045, USA

YUHANG XU*

Department of Applied Statistics and Operations Research, Bowling Green State University, Maurer Center, Ridge St, Bowling Green, OH 43403, USA
xuy@bgsu.edu

YUMOU QIU

Department of Statistics, Iowa State University, 3214 Snedecor, 2438 Osborn Dr Ames, IA 50011, USA

ROBERT POWERS*

Department of Chemistry, University of Nebraska - Lincoln, 639 N. 12th Street, Lincoln, NE 68588, USA
rpowers3@unl.edu

Summary

The process of identifying and quantifying metabolites in complex mixtures plays a critical role in metabolomics studies to obtain an informative interpretation of underlying biological processes. Manual approaches are time-consuming and heavily reliant on the knowledge and assessment of nuclear magnetic resonance (NMR) experts. We propose a shifting-corrected regularized regression method, which identifies and quantifies metabolites in a mixture automatically. A detailed algorithm is also proposed to implement the proposed method. Using a novel weight function, the proposed method is able to detect and correct peak shifting errors caused by fluctuations in experimental procedures. Simulation studies show that the proposed method performs better with regard to the identification and quantification of metabolites in a complex mixture. We also demonstrate real data applications of our method using experimental and biological NMR mixtures.

Keywords: Chemical shift; NMR metabolomics; Regularized regression; Spectral data.

1. Introduction

Over the last several decades, the field of metabolomics has increasingly gained attention among postgenomics technologies (Dieterle *and others*, 2006) due to its ability to study the state of a biological system at the molecular level. In particular, metabolites are the direct outcomes of all genomic, transcriptomic, and

*To whom correspondence should be addressed.

proteomic responses to environmental stimuli, stress, or genetic mutations (Fiehn, 2002). Small changes in metabolite concentration levels might reveal crucial information that is closely related to disease status (Gowda *and others*, 2008), drug resistance (Thulin *and others*, 2017), and the biological activity of chemicals derived from diet and/or environment (Daviss, 2005). Therefore, metabolomics has become an increasingly attractive approach for researchers in many scientific areas such as toxicology (Ramirez *and others*, 2013), food science and nutrition (Wishart, 2008a), and medicine (Putri *and others*, 2013).

Nuclear magnetic resonance (NMR) spectroscopy is one of the premier analytical platforms to acquire data in metabolomics. It is renowned for the richness of information, rapid and straightforward measurements, high level of reproducibility, and minimal sample preparation (Wishart, 2008b). Each metabolite is uniquely characterized by its own resonance signature, namely ^1H NMR chemical shift fingerprint. Every spectral peak is generated by a distinct hydrogen nucleus resonating at a particular frequency, which is measured in parts per million (ppm) relative to a standard compound (Dona *and others*, 2016). For a particular metabolite, depending on its chemical structure, one or more peaks can show up at specific locations on the chemical shift axis. At the same time, the height of every spectral peak is directly proportional to the concentration of the corresponding metabolite in the mixture.

As an illustration, Figure 1 shows individual ^1H NMR spectra of three metabolites (Figure 1(a)–(c)) under an ideal experimental condition. In each panel, the x-axis denotes the *chemical shift* which is measured in ppm while the y-axis represents the relative *peak intensity* corresponding to each chemical shift. Additionally, whenever a *peak* is mentioned, it is referred to a small symmetrical segment of the spectrum; and the chemical shift corresponds to the center of the peak is known as a *peak location*. Ideally, given a mixture spectrum composed of several metabolites as shown in Figure 1(d), one could overlay the figure with each individual reference spectrum such as Figure 1(a)–(c)) to potentially identify each metabolite in the mixture if the signals match. The process of identifying individual metabolites in a complex mixture is called *identification*. Simultaneously, how much each metabolite contributes to the mixture is quantified by their corresponding peak intensities in the mixture spectrum. The process of estimating the concentration of each metabolite in the mixture is called *quantification*. Therefore, the NMR fingerprint and corresponding peak intensities are keys to any approaches to identify and quantify metabolites present in complex biological mixtures.

A conventional approach, which involves manual assignment protocols, has been previously reported (Dona *and others*, 2016). The manual approach relies on experienced spectroscopists to overlay the observed spectrum with reference spectra of pure compounds to decide which particular metabolites are present in the mixtures, so the whole process is time-consuming, labor-intensive, and prone to biases towards operator knowledge and expectations (Tulpan *and others*, 2011). Automating the process of metabolite identification and quantification is desired, but there exists two major obstacles. First, uncontrollable sample perturbations are inherent to every metabolomics study, which arise from a variety of sources such as variation in experimental factors (e.g., pH, temperature, and ionic strength), instrument instability, and inconsistency in sample handling and preparation. As a result, NMR signals of a metabolite may deviate from their referenced positions, which, in turn, makes it harder for any matching procedures. Figure 1(e) illustrates such shifting errors in signal positions, where the glycine peak is shifted to the right of its referenced location at 3.54 ppm (i.e., dashed line). Second, the number of candidate metabolites in the database always exceeds the number of actual sources of signals in the spectra, which raises a sparsity issue. For example, the number of metabolites detected from intact serum/plasma is in the range of 30 or less which is far fewer than the 4229 blood metabolites in the Human Serum Metabolome (Psychogios *and others*, 2011). The combination of the two factors makes the detection and interpretation of metabolite-specific signals challenging in practice.

Regularized regression approaches such as Lasso, elastic net, and adaptive Lasso seem to be intuitive choices to handle the sparsity problem because of their built-in regularization capability. However, they are not capable of addressing peak shifting errors. Recently, high-dimensional regression with measurement

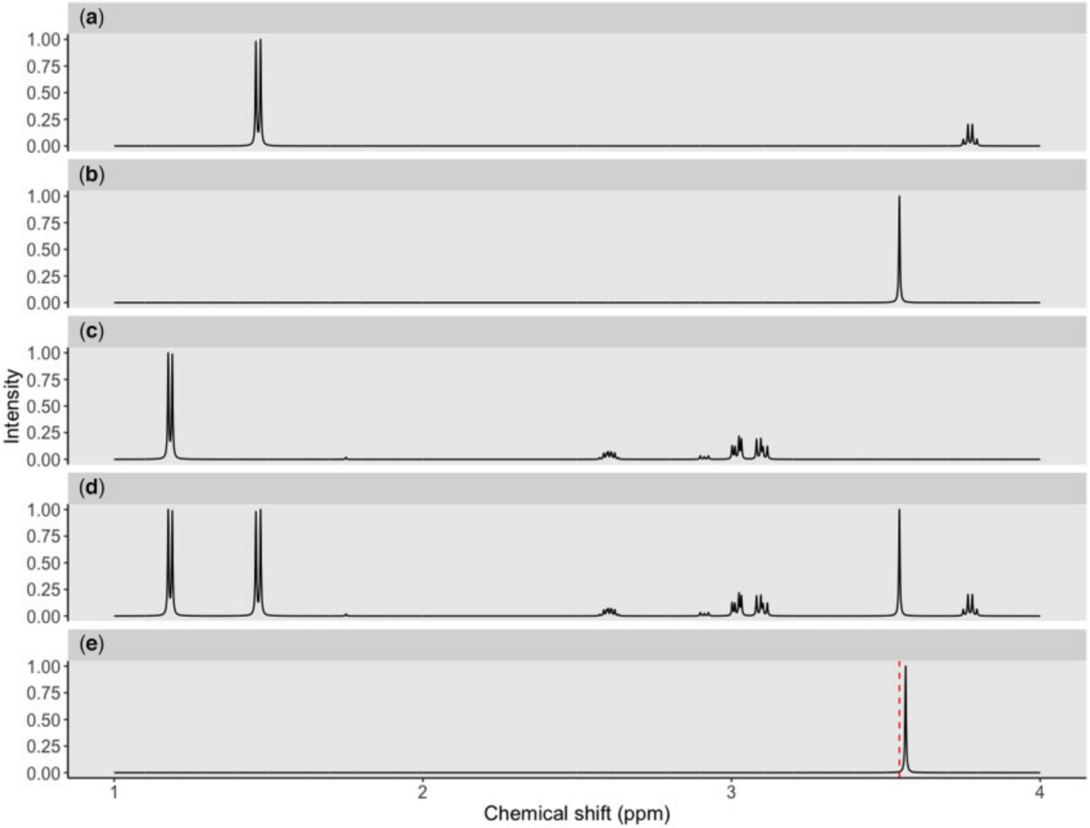

Fig. 1. Three reference spectra of L-alanine (a), glycine (b), and 3-aminoisobutanoic acid (c) convolve a mixture spectrum (d) in an ideal condition. Another mixture spectrum (e) has the glycine peak shifted to the right from the referenced location (dashed line).

errors in covariates is an emerging statistical research area. For example, to deal with the measurement error problem, Sørensen *and others* (2015) and Datta and Zou (2017) proposed different modifications of Lasso; and Sørensen *and others* (2018) introduced methods based on the matrix uncertainty selector (Rosenbaum and Tsybakov, 2010). However, these approaches could not be applied to the problem of shifting errors due to two key differences. First, these works assume that the responses and covariates are correctly matched, but the covariates are subject to additive measurement errors. However, in the discussed problem, the observed spectral intensities of a mixture are assumed to be generated from mismatched covariates, i.e., the intensities of compounds with shifting errors. Second, replicates of covariate measurements or an external validation sample are traditionally required to calibrate the models to deal with measurement errors in covariates (Carroll *and others*, 2006). However, neither of them is available for the type of NMR data being considered. In a different manner, Bayesil (Ravanbakhsh *and others*, 2015), Chenomx (Chenomx, 2015), and ASICS (Tardivel *and others*, 2017; Lefort *and others*, 2019) develop their own methodology to deal with both problems. More precisely, Bayesil partitions the sample spectrum into disjoint regions before applying a probabilistic approach to assign a low probability to an undesirable match and vice versa. Additionally, an automated Profiler module of a popular proprietary software, namely Chenomx, utilizes a linear combination of Lorentzian peak shape models of reference metabolites to reconstruct the observed mixture spectrum (Weljie *and others*, 2006). Uniquely, ASICS learns warping

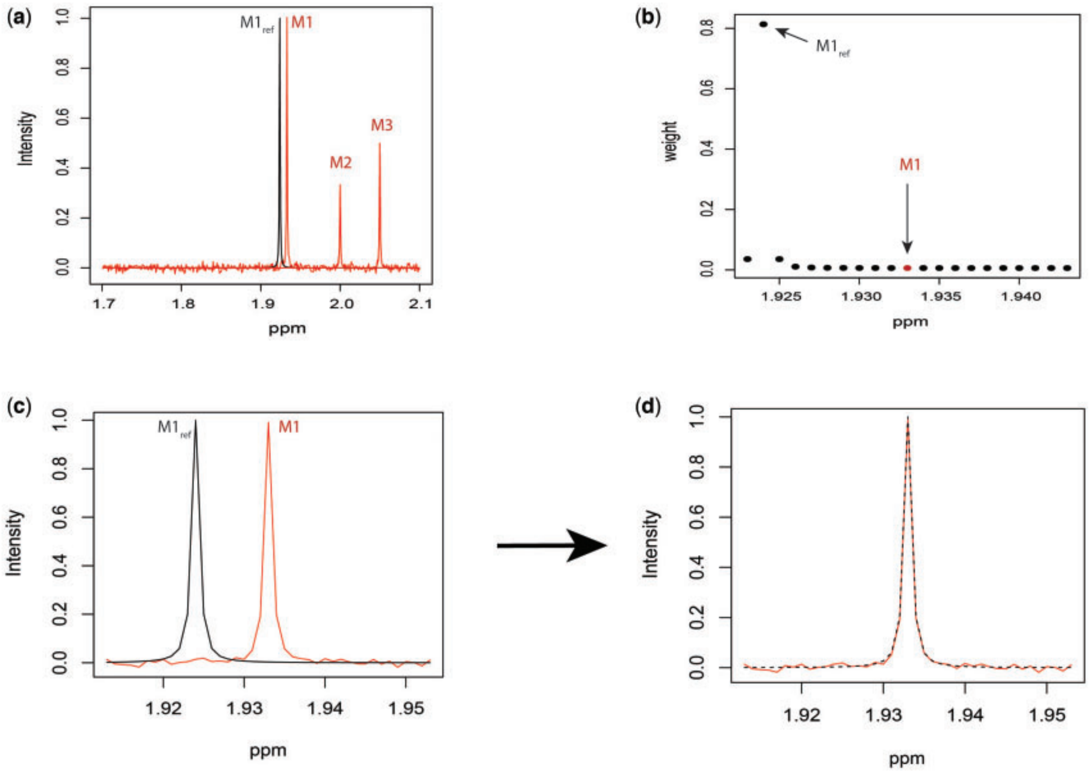

Fig. 2. Simulated mixture spectrum (three rightmost peaks) with added random noise (a) overlaid with the reference counterpart $M1_{\text{ref}}$ (leftmost peak) of $M1$ in the simulation. Weight plot for the shifted peak (b) potentially relocates the shifted peak by detecting the noncentered maximum weight. As a result, shifted peak (c) is corrected to match its referenced peak (d).

functions to minimize the difference between the observed and reconstructed spectra before quantifying individual metabolite concentration. However, none of the methods has yet been demonstrated to be a gold standard in practice.

Herein, we introduce a new approach to automatically identify and quantify metabolites in complex biological mixtures. This parsimonious proposed method is shown to be efficient by simultaneously addressing both problems of shifting errors and the sparsity of some abundant metabolites present in mixtures. Specifically, the method first conducts the variable selection to identify correct metabolites in a mixture with nonzero coefficients. Second, the method performs a postselection coefficient estimation to quantify metabolite concentration after correcting for shifting errors using an embedded novel weight function. We demonstrate the effectiveness of the proposed model using simulated data, experimental NMR mixtures, and biological serum samples. Interesting findings are further emphasized when the method is compared with popular regularized regression models including Lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), and adaptive Lasso (Zou, 2006), and other existing fitting models including Bayesil, Chenomx, and ASICS.

2. Model and methodology

2.1. *Backgrounds*

Each NMR spectrum after being preprocessed by apodization, phasing, and baseline correction can be represented as a pair of equally spaced vector of chemical shifts typically ranging from 0 to 10 ppm and a same length vector of the corresponding relative intensity of the resonance. Depending on the resolution of the instrument or the spectrum, the total number of features in a spectrum is in the order of 10^3–10^4 (Astle *and others*, 2012). However, some NMR signals with low intensities might correspond with instrumental noise, which are not reliable for identifying metabolites in a complex mixture. Thus, we define each NMR spectrum of interest as a pair of $(\mathbf{x}, \mathbf{y})$, where $\mathbf{y} = \{y_i\}_{i=1}^n$ are the observed collection of signal intensities such that $\forall y_i > c_0$, and $\mathbf{x} = \{x_i\}_{i=1}^n$ are the corresponding chemical shifts. Here, a positive constant c_0 serves as a threshold to remove low-intensity signals that are likely to be noise while reducing the number of features to be considered in our model. Details about the selection of c_0 are described in Section 4.

2.2. *Spectrum model with shifting errors*

A major underlying assumption in NMR-based quantitative metabolomics is that any given mixture spectrum is the accumulated sum of individual metabolite spectra (Wishart, 2008b). As illustrated in Figure 1, the peaks of the mixture in Figure 1(d) are composed of the three spectra in Figure 1(a)–(c). In this regard, the abundance of an individual metabolite is reflected by its relative peak heights. Consequently, a spectral representation of a mixture consisting of individual metabolites can be considered as a linear combination of spectral functions of each individual metabolite in the reference library. At a chemical shift x_i, its corresponding intensity of a true mixture spectrum in an ideal experimental condition, denoted by $y_i^\dagger$ can be modeled as follows:

$$y_i^\dagger = \beta_0 + \sum_{j=1}^{p} \beta_j g_j(x_i) + \epsilon_i, \tag{2.1}$$

where $i = 1, \ldots, n$ are all indices of chemical shifts of the peaks along the mixture spectrum; p is the number of known compounds in the reference library; $g_j(x_i), j = 1, \ldots, p$ is the intensity function of the jth reference spectrum; ϵ_i represents random noise with mean zero and variance σ_ϵ^2; and non-negative β_j represents the concentration of the jth metabolite in the complex mixture. Accordingly, the jth metabolite is considered to be present in the mixture if the coefficient β_j is greater than 0. By mean-centering $y_i^\dagger$ and $g_j(x_i)$ such that $\sum_{i=1}^n y_i^\dagger = 0$ and $\sum_{i=1}^n g_j(x_i) = 0$, we can remove the intercept term β_0 from (2.1) (Tibshirani, 1996).

Each reference spectrum is considered as a collection of peaks with different chemical shift locations and peak intensities. Since NMR peaks are sharp, it is common to represent each NMR peak as a Lorentzian curve (i.e., Cauchy distribution function) (Hollas, 2004). Depending on the molecular environment and the size of the molecule, the number of peaks in a ^1H NMR spectrum can range from 1 (e.g., methanol) to more than 47 (e.g., D-glucose). For the jth metabolite with n_j ($n_j \geq 1$) chemical shift positions in the reference library, its spectrum can be modeled as follows:

$$g_j(x; \mathbf{l}_j, \mathbf{r}_j) = \sum_{m=1}^{n_j} v_{jm} \frac{1}{1 + (\frac{x - l_{jm}}{r_{jm}})^2}, \tag{2.2}$$

where x is an input which can take any value along the chemical shift (ppm) axis; n_j is a total number of peaks of the jth metabolite; $\mathbf{l}_j = (l_{j1}, \ldots, l_{jn_j})$ is a vector of all peak locations of the jth metabolite; $\mathbf{r}_j = (r_{j1}, \ldots, r_{jn_j})$ is a vector of shape parameters for each of the n_j peaks, and these values are set at 0.002 to maintain the sharp shape of an NMR peak (Vu *and others*, 2019). For notation simplicity, we remove $\mathbf{r}_j$ from $g_j(x; \mathbf{l}_j, \mathbf{r}_j)$ for the rest of the paper. Finally, $\mathbf{v}_j = (v_{j1}, \ldots, v_{jn_j})$ is the multiplier factor for each of the n_j peaks such that the relative ratios between peak heights are maintained. For each metabolite, we obtain a list of peak locations and corresponding relative peak heights directly from the Human Metabolome

Database (Wishart *and others*, 2018). Here, a vector of multiplier factor $\mathbf{v}_j$ is calculated by solving linear equations of Cauchy densities evaluated at each peak location and corresponding peak heights; see Vu *and others* (2019) for details. Using reference spectra generated directly from (2.2) has the advantage over in-house spectra of pure chemical compounds in terms of minimizing some undesirable experimental perturbations. From (2.1), the peaks of the reference spectra with $\beta_j > 0$ should also be peaks among $\{y_i^\dagger\}$ of the target mixture. However, unavoidable fluctuations in sample pH, temperature, and instrument instability can cause peaks of the mixture to shift from their referenced locations. As a result, the observed spectrum intensities may incur location shifting errors. In this regard, the observed intensity y_i at x_i is subject to a location shift such that

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j g_j(x_i; \mathbf{l}_j + \boldsymbol{\delta}_j) + \epsilon_i, \tag{2.3}$$

where $\boldsymbol{\delta}_j = (\delta_{j1}, \ldots, \delta_{jn_j})$ is a vector of shifting errors associated with referenced peak locations $\mathbf{l}_j$. In other words, when a particular peak is shifted, the neighboring signals are accordingly shifted by the same amount. Each $\{\delta_{jm}\}_{m=1}^{n_j}$ follows a distribution $F(\cdot)$ with a bounded support on $[-K, K]$ for a positive constant K. The bounded support ensures the locality of shifting errors associated with signals in the mixture spectrum. For a given reference spectrum, we know the parameters $\mathbf{l}_j$ and $\mathbf{v}_j$ in (2.2). However, the shifting errors $\boldsymbol{\delta}_j$ are not observable. Hence, a direct regression of y_i on $\{g_j(x_i; \mathbf{l}_j + \boldsymbol{\delta}_j)\}_{j=1}^{p}$ to estimate $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ based on the model (2.3) is not practical.

Models (2.1) and (2.3) imply a mismatch between the observed and referenced intensities (i.e., y_i and $y_i^\dagger$) of the mixture. Such shifting deviations need to be corrected to ensure the consistent estimation of $\boldsymbol{\beta}$ and accurate identification of the compounds present in the mixture. In Section 2.3, we propose a shifting-corrected regularized regression estimation procedure to correct for the positional errors in the spectral signals.

2.3. *Methodology*

The total number p of metabolites in the reference library used for spectral fitting is typically in the order of 10^2–10^3 depending on the types of sample mixtures. The number of abundant metabolites actually present in the mixtures is a small subset of the reference library. Namely, most of the coefficients of the compounds not contributing to the mixture should ideally be zero in the regression model (2.1). Given the sparsity feature of the problem, we apply the Least Absolute Shrinkage and Selection Operator (Lasso) regularization to obtain a sparse estimate of the regression coefficients (Tibshirani, 1996). Recall that $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$. If the spectra intensities $\{y_i^\dagger\}$ without shifting errors can be observed, we can estimate $\boldsymbol{\beta}$ by minimizing the following objective function

$$L^\dagger(\boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^{n} \left\{ y_i^\dagger - \beta_0 - \sum_{j=1}^{p} \beta_j g_j(x_i; \mathbf{l}_j) \right\}^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.4}$$

where λ is a penalty parameter. With proper selection of λ, Lasso is capable of obtaining sparse estimate that is consistent to $\boldsymbol{\beta}$ (Bickel *and others*, 2009; Bühlmann and Van de Geer, 2011). When $\lambda = 0$, i.e., no penalty is applied, Lasso-type estimates are simply ordinary least square estimates. As λ increases, more β_j are shrunk to exactly zero (James *and others*, 2013). However, (2.4) cannot be implemented as the intensities $\{y_i^\dagger\}$ without shifting errors are not observable.

Let $s_{ik} = \{y_i - \beta_0 - \sum_{j=1}^{p} \beta_j g_j(x_k; \mathbf{l}_j)\}^2$ be the squared distance between the observed signal intensity at x_i and the reconstructed intensity from reference spectra at x_k. If the observed peak at x_i is in fact generated from $\sum_{j=1}^{p} \beta_j g_j(x_k; \mathbf{l}_j)$, i.e., $y_i = y_k^{\dagger}$, there exists a shifting error in signal locations of $\delta_0 = x_k - x_i$. Then, the residual term s_{ik} should be small. Otherwise, the value of s_{ik} should be relatively large. For each $i = 1, \ldots, n$, we calculate such pairwise residuals for the interval $\{\min(1, i-d), \ldots, i, \ldots, \max(i+d, n)\}$, where d is a predefined, positive constant. In practice, d may be empirically chosen and details about its selection are discussed in Section 4. The residuals s_{ik} can be used to construct weights for each feature pair (i, k). Let $\phi(z; \sigma_0) = \exp\{-z^2/(2\sigma_0^2)\}$ be the kernel of the normal density function with mean $\mu = 0$ and variance σ_0^2. Define the weight function as follows:

$$w_{ik}(\boldsymbol{\beta}) = \frac{\phi(s_{ik}; \sigma_0)}{\sum_{k=k_l}^{k_u} \phi(s_{ik}; \sigma_0)}, \tag{2.5}$$

where $k_l = \min(1, i-d), k_u = \max(i+d, n)$; σ_0 serves as a tuning parameter which controls the distribution of weights in each search window $\{\min(1, i-d), \ldots, i, \ldots, \max(i+d, n)\}$. For simplicity, we will use k_l and k_u as defined above for the rest of the article. Notice that w_{ik} is a smooth and decreasing function of s_{ik} with $\sum_{k=k_l}^{k_u} w_{ik}(\boldsymbol{\beta}) = 1$ for each $i = 1, \ldots, n$. For each search window $\{\min(1, i-d), \ldots, i, \ldots, \max(i+d, n)\}$, the weight reaches its maximum if the observed signal at x_i is shifted from its reference counterparts at x_k and hence s_{ik} is the smallest. For the rest of the article, w_{ik} will be used in place of $w_{ik}(\boldsymbol{\beta})$ to simplify the notation.

Using the weights in (2.5), we propose a shifting-weighted regularized estimation approach that minimizes the following objective function:

$$\begin{aligned} L(\boldsymbol{\beta}) &= \frac{1}{2} W(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j \ \text{ subject to } \ \beta_j \geq 0 \ \text{ for all } j = 1, \ldots, p, \text{ where} \\ W(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \sum_{k=k_l}^{k_u} w_{ik} \left\{ y_i - \beta_0 - \sum_{j=1}^{p} \beta_j g_j(x_k; \mathbf{l}_j) \right\}^2. \end{aligned} \tag{2.6}$$

Here, (2.6) is a constrained regularized optimization, where the non-negativity constraint on β_j is due to the non-negativity of metabolite concentrations in our problem. For general problems without constraints, we may impose the penalty $\lambda \sum_{j=1}^{p} |\beta_j|$ in (2.6). Note that this optimization problem is more complex than the classical Lasso optimization and may not be convex, since the weights w_{ik} also depend on the regression coefficients $\boldsymbol{\beta}$.

Compared to (2.4), the loss function $W(\boldsymbol{\beta})$ takes into account any potential signal shifting for each location x_i by including the pairwise distance s_{ik} corresponding to each element in the search window $\{\min(1, i-d), \ldots, i, \ldots, \max(i+d, n)\}$. These pairwise distances are weighted by w_{ik} such that the large s_{ik} is multiplied by a small value and vice versa, where the weight w_{ik} decays exponentially as s_{ik} increases.

Let $\hat{\boldsymbol{\beta}}$ be a minimizer of (2.6) where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^{\mathrm{T}}$. Additionally, let $\mathcal{A} = \{j : j \in \{1, \ldots, p\}, \hat{\beta}_j > 0\}$ be the active set of the metabolites which are identified as present in the target mixture. We define $\hat{s}_{ik} = \{y_i - \hat{\beta}_0 - \sum_{j \in \mathcal{A}} \hat{\beta}_j g_j(x_k; l_j, r_j)\}^2$, and $\hat{w}_{ik} = \phi(\hat{s}_{ik}; \sigma_0)\big(\sum_{k=k_l}^{k_u} \phi(\hat{s}_{ik}; \sigma_0)\big)^{-1}$. At each peak location x_i of the mixture, the value $\arg\max_k\{\hat{w}_{ik}\}$ together with the reference peak locations around x_i can be used to estimate and correct for the shifting errors.

The tuning parameter σ_0 can be considered as a weight distributor for each search window $\{\min(1, i-d), \ldots, i, \ldots, \max(i+d, n)\}$ corresponding to x_i. Smaller σ_0 yields a narrower weight distribution, which results in more weights close to 0. In this regard, an extremely small σ_0 would assign the weight of 1 to the smallest s_{ik} while the remaining weights are essentially 0. On the other hand, a large σ_0 would flatten out the weight distribution, which in turn loses the ability to detect the signal shifting. Given $g_j(x_k; \mathbf{l}_j) > 0$ $\forall j, k, s_{ik}$ takes a value between 0 and y_i^2. In general, we suggest σ_0 to be between $\max(y_i^2)/3$ and $\max(y_i^2)/6$,

$i = 1, \ldots, n$, to maintain the smoothness in weight distribution. More discussion about the sensitivity of σ_0 on the performance of the proposed method is included in Section 4.

3. Implementation

In this section, we provide the computation algorithms to solve the shifting-corrected regularized estimation (2.6) proposed in Section 2.3.

3.1. *Coordinate descent*

As both w_{ik} and $\{y_i - \beta_0 - \sum_{j=1}^{p} \beta_j g_j(x_k; \mathbf{l}_j)\}^2$ in the objective function $L(\boldsymbol{\beta})$ in (2.6) depend on $\boldsymbol{\beta}$, it might not be a convex function of $\boldsymbol{\beta}$. However, for any fixed positive weights $\{w_{ik}\}$, $W(\boldsymbol{\beta})$ is a weighted least squares loss of the augmented paired data $\{(y_i, x_k)_{k=k_l}^{k_u}\}_{i=1}^{n}$; hence, it is a convex function. Therefore, to minimize $L(\boldsymbol{\beta})$, we utilize the coordinate descent approach (Friedman *and others*, 2010). This optimization process minimizes the objective function with respect to each β_j at a time while fixing the weights w_{ik} and the remaining coefficients $\{\beta_h\}_{h \neq j}$. Specifically, the gradient of $W(\boldsymbol{\beta})$ with respect to β_j, given fixed weights w_{ik}, is $\frac{\partial}{\partial \beta_j} W(\beta) = -(\rho_j - \beta_j z_j)$, where $\rho_j = \sum_{i=1}^{n} \sum_{k=k_l}^{k_u} w_{ik} \{y_i - \beta_0 - \sum_{h \neq j}^{p} \beta_h g_h(x_k; \mathbf{l}_h)\} g_j(x_k; \mathbf{l}_j)$, and $z_j = \sum_{i=1}^{n} \sum_{k=k_l}^{k_u} w_{ik} \{g_j(x_k; \mathbf{l}_j)\}^2$. Details of the derivation are provided in Section S1 of the Supplementary material available at *Biostatistics* online. Since the penalty term $\lambda \sum_{j=1}^{p} \beta_j$ in (2.6) is separable in $\boldsymbol{\beta}$, for each component j, $L(\boldsymbol{\beta})$ can be expressed as

$$L(\beta_j; \boldsymbol{\beta}_{-j}) = \sum_{i=1}^{n} \sum_{k=k_l}^{k_u} w_{ik} \{y_i - \beta_j g_j(x_k; \mathbf{l}_j) - C_1(\beta_0, \boldsymbol{\beta}_{-j})\}^2 + \lambda \beta_j + C_2(\boldsymbol{\beta}_{-j}),$$

where $C_1(\beta_0, \boldsymbol{\beta}_{-j})$ and $C_2(\boldsymbol{\beta}_{-j})$ are two functions independent of β_j, and $\boldsymbol{\beta}_{-j}$ denotes the regression coefficients without the jth component. Therefore, the objective function $L(\boldsymbol{\beta})$ is a quadratic convex function of β_j given all other coefficients. The coordinate descent algorithm essentially minimizes a quadratic convex function $L(\beta_j; \boldsymbol{\beta}_{-j})$ of β_j with the constraint $\beta_j \geq 0$. Since $\frac{\partial}{\partial \beta_j} L(\beta_j; \boldsymbol{\beta}_{-j}) = \frac{\partial}{\partial \beta_j} W(\beta) + \lambda = \beta_j z_j - \rho_j + \lambda$, given $\boldsymbol{\beta}_{-j}$ and β_0 fixed, the minimum of $L(\beta_j; \boldsymbol{\beta}_{-j})$ over $\beta_j \geq 0$ occurs at $\max\{0, (\rho_j - \lambda)/z_j\}$.

Specifically, at the current estimate $\hat{\boldsymbol{\beta}}^{(u)}$, we obtain the weight functions as

$$\hat{w}_{ik}^{(u)} = \frac{\phi(\hat{s}_{ik}^{(u)}; \sigma_0)}{\sum_{k=k_l}^{k_u} \phi(\hat{s}_{ik}^{(u)}; \sigma_0)}, \tag{3.7}$$

where $\hat{s}_{ik}^{(u)} = \left\{y_i - \hat{\beta}_0^{(u+1)} - \sum_{h_1=1}^{j-1} \hat{\beta}_{h_1}^{(u+1)} g_{h_1}(x_k; \mathbf{l}_{h_1}) - \sum_{h_2=j}^{p} \hat{\beta}_{h_2}^{(u)} g_{h_2}(x_k; \mathbf{l}_{h_2})\right\}^2$. Sequentially, we obtain

$$\rho_j^{(u)} = \sum_{i=1}^{n} \sum_{k=k_l}^{k_u} \hat{w}_{ik}^{(u)} \left\{ y_i - \hat{\beta}_0^{(u+1)} - \sum_{h_1=1}^{j-1} \hat{\beta}_{h_1}^{(u+1)} g_{h_1}(x_k; \mathbf{l}_{h_1}) - \sum_{h_2=j+1}^{p} \hat{\beta}_{h_2}^{(u)} g_{h_2}(x_k; \mathbf{l}_{h_2}) \right\} g_j(x_k; \mathbf{l}_j) \quad \text{and} \tag{3.8}$$

$$z_j^{(u)} = \sum_{i=1}^{n} \sum_{k=k_l}^{k_u} \hat{w}_{ik}^{(u)} \{g_j(x_k; \mathbf{l}_j)\}^2.$$

Note that at each uth iteration the process is done for all β_j's ($j = 1, \ldots, p$). Then, we obtain the $(u+1)$th update of β_j by

$$\hat{\beta}_j^{(u+1)} = \max\{0, (\rho_j^{(u)} - \lambda)/z_j^{(u)}\}. \tag{3.9}$$

It is worth noting that if there is no non-negativity constraint on β_j for all j, and the Lasso penalty $\lambda \sum_{j=1}^{p} |\beta_j|$ is used in (2.6), the coordinate descent algorithm updates β_j by the soft-thresholding operator as done by Friedman *and others* (2007).

At each iteration, the algorithm updates each regression coefficient β_j separately, which requires $O(p)$ computation steps. Meanwhile, there are $n(2d+1)$ weights to update for each updated β_j. The total computational complexity is $O\{np(2d+1)\}$ per iteration. Additionally, the process of looping through all regression coefficients β_j is iterated until the convergence criterion $\|\hat{\boldsymbol{\beta}}^{(u)} - \hat{\boldsymbol{\beta}}^{(u-1)}\| < 10^{-5}$ is met or when the maximum number of iterations, which is set at 1000, is reached. Intuitively, as $W(\boldsymbol{\beta})$ is a convex function given the weights $\{w_{ik}\}$, the algorithm would converge if the initial weights are close to the ones with the true $\boldsymbol{\beta}$. While $W(\boldsymbol{\beta})$ is a nonconvex function of β as $\{w_{ik}\}$ changes with β, and our proposed algorithm is not guaranteed to converge to a global optimum, we find that the results are not sensitive to the initial values in the simulation studies and the real data analysis. The theoretical convergence properties of the proposed method will be investigated in future work.

Let $\mathbf{G}_{n\times p}$ be the reference library data matrix, where the jth column of $\mathbf{G}$ consists of the spectrum $\{g_j(x_i; \mathbf{l}_j)\}_{i=1}^{n}$ of the jth metabolite in (2.2). As before, $\mathbf{y}_{n\times 1}$ is the n-dimensional vector representing the spectrum of the target mixture. Given the penalty parameter λ chosen by the cross-validation (CV) criterion, the tuning parameter σ_0 in the weight function (2.5), and the search window size d, the main steps of the proposed optimization algorithm are outlined below.

Algorithm 1 Coordinate descent algorithm to solve $\boldsymbol{\beta}$ in (2.6)

```
Standardize each column of $\mathbf{G}_{n\times p}$ as $\mathbf{G}_{c,ij} = \frac{\mathbf{G}_{ij}-\bar{\mathbf{G}}_j}{\mathrm{sd}_j}$ where $\mathrm{sd}_j$ is the standard deviation of the $j$th column of $\mathbf{G}$; center $\mathbf{y}$ to have $\mathbf{y}_c = \mathbf{y} - \bar{\mathbf{y}}$;
$u \leftarrow 0$; initialize $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}}^{(0)} = 0_{p\times 1}$;
while ($\|\hat{\boldsymbol{\beta}}^{(u)} - \hat{\boldsymbol{\beta}}^{(u-1)}\| \geq 10^{-5}$ or # iterations $\leq 1000$) do
    for $j = 1, 2 \ldots, p$ do
        obtain weight function $\hat{w}_{ik}^{(u)}$ as in (3.7);
        compute $\rho_j^{(u)}$ and $z_j^{(u)}$ using (3.8);
        update $\hat{\beta}_j^{(u+1)}$ based on (3.9);
    end for
    $u \leftarrow u + 1$
end while
return $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1^{(u)}, \ldots, \hat{\beta}_p^{(u)})$ as the first-stage estimated regression coefficients.
```

3.2. *Cross-validation*

A decreasing sequence of λ values is used to calculate the corresponding prediction errors through CV. The optimal penalty parameter λ associated with the smallest error is chosen. Similar to Lasso's pathwise coordinate descent (Friedman *and others*, 2010), the sequence of λ is generated such that its maximum ($\lambda_{\max}$) is the minimum penalty value that all the estimated coefficients become 0. Specifically, $\lambda_{\max}$ is computed as $\lambda_{\max} \geq \left|\sum_{i=1}^{n} w_{ik}^{*} y_i g_j(x_k; \mathbf{l}_j)\right|$ where $w_{ik}^{*} = \frac{\phi(s_{ik}^{*})}{\sum_k \phi(s_{ik}^{*})} = \frac{1}{2d+1}$ with $s_{ik}^{*} = y_i^2$. Then, we define $\lambda_{\min} = c\lambda_{\max}$ for a small positive value of c. The λ sequence of length M_0 is constructed by linearly decreasing from $\lambda_{\max}$ to $\lambda_{\min}$ on a log scale, where c and M_0 are recommended to be 0.0001 and 50, respectively according to Friedman *and others* (2010).

The constructed sequence of penalty values λ is then used for 5-fold CV as outlined in Algorithm 2. For both the observed mixture spectrum (response) and the spectra in the reference library (covariates), all NMR signals after thresholding with c_0 are randomly partitioned into five sets. During the random partition, it is almost certain that some signals of each peak are used for training, which is sufficient to detect any peak shifting. Each of the five sets is used for validation while the other four sets are used for training. For a given value of λ in the sequence, the regression coefficients are estimated using the training set. The estimated loss $W(\boldsymbol{\beta})$ in (2.6), which is obtained using the estimated coefficients, is evaluated on the validation set. The CV loss corresponding to each λ is the average loss over the five sets; the cross-validated penalty value is chosen as the minimizer of the CV loss. Similarly, we apply the standard 5-fold CV to choose the penalty values for Lasso, elastic net, and adaptive Lasso. The main difference is that the three methods utilize standard least-squares loss while the proposed method uses the weighted loss in (2.6) to account for the shifting errors.

Algorithm 2 Cross-validation

```
Create a decreasing lambda sequence $\lambda_{\text{seq}}$ using $\lambda_{\min}$, $\lambda_{\max}$, and $M_0$ as in Section 3.2
Initialize $k_0$ folds;
Randomly partition $n$ indices of $\mathbf{y}$ and $\mathbf{G}$ into $k_0$ folds;
for $\lambda_j$ in $\lambda_{\text{seq}}$ do
    for fold$_i$ in the $k_0$ folds do
        $\mathbf{y}_{\text{test}} = \mathbf{y}[\text{fold}_i]$; $\mathbf{G}_{\text{test}} = \mathbf{G}[\text{fold}_i, ]$; $\mathbf{y}_{\text{train}} = \mathbf{y}[-\text{fold}_i]$; $\mathbf{G}_{\text{train}} = \mathbf{G}[-\text{fold}_i, ]$;
        $\hat{\boldsymbol{\beta}}_{\text{fold}_i} =$ Algorithm 1$(\mathbf{y}_{\text{train}}, \mathbf{G}_{\text{train}}, \lambda_j)$;
        calculate the loss function $W(\hat{\boldsymbol{\beta}}_{\text{fold}_i})$ evaluated on $\mathbf{y}_{\text{test}}$ and $\mathbf{G}_{\text{test}}$ according to (2.6);
    end for
    obtain the average loss $W(\hat{\boldsymbol{\beta}})_{\lambda_j}$ across all fold$_i$;
end for
$\lambda_{\text{optimal}} = \arg\min_{\lambda_j} W(\hat{\boldsymbol{\beta}})_{\lambda_j}$.
```

3.3. *Concentration estimation*

Let $\hat{\boldsymbol{\beta}}$ be the solution of the coordinate descent procedure in Section 3.1, and $\{\hat{w}_{ik}\}$ be the corresponding estimated weights. To estimate the concentration of the present metabolites, we first need to correct for the shifting errors. At each signal location x_i, let $\hat{x}_i^{(\max)} = \arg\max_k\{\hat{w}_{ik}\}$ be the position corresponding to the smallest pairwise distance $\hat{s}_{ik}$ for $i = 1, \ldots, n$. For each jth metabolite with $\hat{\beta}_j > 0$, we match $\hat{x}_i^{(\max)}$ with its referenced peak locations $\{l_{jm}\}_{m=1}^{n_j}$. The shifting error δ_{jm} associated with the mth peak of the jth reference metabolite at x_i can be estimated as

$$\hat{\delta}_{jm} = \begin{cases} x_i - l_{jm}, & \text{if } l_{jm} = \hat{x}_i^{(\max)} \\ 0, & \text{if } l_{jm} \neq \hat{x}_i^{(\max)} \end{cases}$$

for $i = 1, \ldots, n$, and all $j \in \mathcal{A}$, where $\mathcal{A} = \{j : j \in \{1, \ldots, p\}, \hat{\beta}_j > 0\}$ is the active set of the metabolites identified in the target mixture. Let $\hat{\boldsymbol{\delta}}_j = \{\hat{\delta}_{jm}\}_{m=1}^{n_j}$. The final estimation of the metabolites concentration after the adjustment to shifting errors is denoted by $\tilde{\boldsymbol{\beta}}$, where $\tilde{\beta}_j = 0$ if $j \notin \mathcal{A}$, and $\{\tilde{\beta}_j\}_{j\in\mathcal{A}}$ can be obtained by minimizing the following objective function directly

$$\sum_{i=1}^{n}\left\{y_i - \sum_{j\in\mathcal{A}}\beta_j g_j(x_i;\mathbf{l}_j + \hat{\boldsymbol{\delta}}_j)\right\}^2. \tag{3.10}$$

Here, the non-negativity constraint is again enforced such that $\tilde{\beta}_j = 0$ for $j \in \mathcal{A}$ if $\tilde{\beta}_j < 0$ to ease the interpretation of non-negative metabolite concentration. Additionally, as the concentration estimation $\tilde{\boldsymbol{\beta}}$ in (3.10) is conditional on the selection results, i.e., the estimation of $\hat{\boldsymbol{\beta}}$ in (3.9) as well as the correction for shifting error, the inference for $\tilde{\boldsymbol{\beta}}$ is more complicated than the usual post-selection Lasso estimators. In order to study the impact of the two steps on the least square estimator $\tilde{\boldsymbol{\beta}}$, one could consider the stability selection procedure, as in Meinshausen and Bühlmann (2010). More discussion about this is described in Section S1.3 of the Supplementary material available at *Biostatistics* online.

4. Simulation studies

The evaluation criteria used to compare the performance of different methods were accuracy, sensitivity, and specificity. Accuracy was calculated as a ratio of correctly labeled metabolites (true positives plus true negatives) to the total number of metabolites in reference library. Similarly, sensitivity was obtained as a fraction of the correctly identified metabolites (true positives) relative to the total number of true metabolites. Moreover, specificity was measured by dividing the number of unidentified metabolites by the number of metabolites not present in a mixture. The correct or incorrect metabolites in a mixture were determined based on their corresponding postselection error-corrected least squares estimates $\tilde{\boldsymbol{\beta}}$ defined in (3.10). Additionally, Figure S1 of the Supplementary material available at *Biostatistics* online showed a decline in the loss function as the number of iterations increased. More precisely, the objective function was stabilized within the first 50 iterations across simulated and real mixtures, which verified the convergence of the proposed algorithm in practice.

Two simulation studies were conducted with a reference database of 200 compounds generated directly from (2.2) for chemical shifts ranging from 0.9 to 9.2 ppm with an equal space of 0.001 ppm, excluding the water suppression region between 4.6 and 4.8 ppm. The peak list for each compound $\mathbf{l}_j = (l_{j1}, \ldots, l_{jn_j})$ was then randomly selected from the available chemical shifts, with n_j ranging from 1 to 10. The corresponding peak heights were generated from the uniform distribution (0.1, 1), which were then used to calculate the multiplier factor $\mathbf{v}_j = (v_{j1}, \ldots, v_{jn_j})$ as described in Section 2.2. The shape parameter r_j was fixed at 0.002 to maintain the sharp shape of an NMR peak. Each resulting spectrum was standardized such that its maximum peak intensity was set to 1. Our simulation studies only consisted of comparisons between the proposed methods and existing regularized regression models (i.e., Lasso, elastic net, and adaptive Lasso) because Bayesil, Chenomx, and ASICS only handled raw ^1H NMR data which were not obtainable through simulation. The performance comparison between the proposed method and existing software including Bayesil, Chenomx, and ASICS, was illustrated in Section 5. Due to limited space, we only reported in detail one of the simulation studies in the main text. The additional simulation study was discussed in Section S2 of the Supplementary material available at *Biostatistics* online.

A target mixture in the simulation as shown in Figure 2(a) was created by adding three individual spectra with random noise to resemble experimental variations. More specifically, true parameters were set up such that $\beta_1 = \beta_2 = \beta_3 = 1$, and $\beta_j = 0$ for $j = 4, \ldots, 200$. Furthermore, positional noise, i.e., peak shifting errors were explicitly examined by purposely shifting locations of chosen peaks. The peak $M1$ in Figure 2(a) was shifted to the right from its referenced location $M1_{ref}$ while $M2$ and $M3$

Table 1. *Average accuracy, sensitivity, and specificity for 200 iterations in the simulation at an increasing shifting variation from* ±0.01 *ppm to* ±0.04 *ppm across the proposed method, Lasso, elastic net, and adaptive Lasso. Corresponding standard deviations are recorded in parentheses*

	Metrics	±0.01ppm	±0.02ppm	±0.03ppm	±0.04ppm
Proposed method	Accuracy	0.999 (0.002)	0.998 (0.003)	0.997 (0.005)	0.995 (0.006)
	Sensitivity	0.980 (0.119)	0.990 (0.081)	0.978 (0.116)	0.950 (0.208)
	Specificity	0.999 (0.002)	0.998 (0.003)	0.997 (0.004)	0.996 (0.005)
Lasso	Accuracy	0.988 (0.029)	0.973 (0.040)	0.956 (0.062)	0.949 (0.065)
	Sensitivity	0.748 (0.328)	0.758 (0.263)	0.760 (0.209)	0.697 (0.270)
	Specificity	0.991 (0.030)	0.976 (0.041)	0.960 (0.063)	0.954 (0.067)
Elastic net	Accuracy	0.941 (0.089)	0.936 (0.064)	0.923 (0.061)	0.898 (0.090)
	Sensitivity	0.992 (0.052)	0.915 (0.146)	0.843 (0.167)	0.827 (0.167)
	Specificity	0.940 (0.091)	0.936 (0.063)	0.924 (0.062)	0.899 (0.092)
Adaptive Lasso	Accuracy	0.984 (0.043)	0.964 (0.056)	0.956 (0.059)	0.945 (0.069)
	Sensitivity	0.812 (0.242)	0.793 (0.210)	0.765 (0.188)	0.757 (0.188)
	Specificity	0.986 (0.044)	0.967 (0.057)	0.959 (0.061)	0.947 (0.070)

stayed unchanged. The amount of chemical shift variation was applied in an increasing fashion, i.e., $\delta_{11} \sim \text{Unif}(-K, K)$, such that $K = \{0.01, 0.02, 0.03, 0.04\}$ ppm respectively to assess the performance of various methods. The whole process of adding random noise to a generated mixture spectrum and shifting peak locations was repeated 200 times. Section S3.2 of the Supplementary material available at *Biostatistics* online discussed in detail how the proposed method behaved as the variance of the added noise increased. Additionally, based on the sensitivity analysis results (Tables S3–S6 of the Supplementary material available at *Biostatistics* online), we set d defined in (2.5) to be the closest integer capturing the maximum shifting variation. In other words, with equal space of 0.001 ppm between chemical shifts, d was set to be 10, 20, 30, and 40 in correspondence with $K = \{0.01, 0.02, 0.03, 0.04\}$ ppm, respectively.

Once a mixture was created, a threshold level c_0 defined in Section 2.1 was obtained such that c_0 was greater than 7% of the area under the mixture spectrum curve (AUC), i.e., $c_0 = 7\%\text{AUC}$ (Ahmed, 2005). An extended simulation study reported in Section S3.1 of the Supplementary material available at *Biostatistics* online assessed how changing c_0 would affect the performance of the proposed method. We evaluated different values of c_0 (i.e., 5%, 7%, 10%, and 12% AUC) in conjunction with different d values (i.e., 5, 10, 15, 20, and 25). Consistent results across c_0 values served as an assurance to continue both simulation studies and real data analysis using $c_0 = 7\%\text{AUC}$. Furthermore, a joint analysis for various values of both σ_0 and d defined in (2.5) was summarized in Section S3.1 of the Supplementary material available at *Biostatistics* online. The results confirmed the choice of $\sigma_0 = \max(y_i^2)/3$ for the analysis.

Table 1 recorded accuracy, sensitivity, and specificity for each method across four increasing levels of positional perturbations, averaged over 200 iterations. As shifting variations increased from ±0.01 to ±0.04 ppm, all accuracy, sensitivity, and specificity decreased across the four methods. However, the decreasing rates were slightly different across different metrics and methods. Specifically, sensitivity had the fastest dropping rate ($\approx$ 3%) compared to accuracy ($\approx$ 0.4%) and specificity ($\approx$ 0.3%) for the proposed method. Lasso particularly had the lowest sensitivity across all levels of shifting errors because of its tendency toward identifying a large number of compounds that were not truly contributing to the mixture.

Table 2. *Average estimated metabolite concentrations for 200 iterations in the simulation at an increasing shifting variations from* ±0.01 *ppm to* ±0.04 *ppm across proposed method, Lasso, elastic net, and adaptive Lasso. Corresponding standard deviations are recorded in parentheses*

	Truth	$\tilde{\boldsymbol{\beta}}$	±0.01ppm	±0.02ppm	±0.03ppm	±0.04ppm
Proposed method	1	$\tilde{\beta}_1$	0.970 (0.106)	0.984 (0.067)	0.968 (0.152)	0.939 (0.228)
	1	$\tilde{\beta}_2$	0.924 (0.218)	0.953 (0.167)	0.948 (0.184)	0.940 (0.213)
	1	$\tilde{\beta}_3$	0.903 (0.274)	0.946 (0.201)	0.946 (0.199)	0.926 (0.251)
Lasso	1	$\tilde{\beta}_1$	0.017 (0.055)	0.008 (0.034)	0.007 (0.040)	0.002 (0.006)
	1	$\tilde{\beta}_2$	0.725 (0.396)	0.850 (0.309)	0.925 (0.210)	0.859 (0.319)
	1	$\tilde{\beta}_3$	0.755 (0.389)	0.867 (0.298)	0.938 (0.194)	0.868 (0.314)
Elastic net	1	$\tilde{\beta}_1$	0.111 (0.223)	0.064 (0.193)	0.060 (0.195)	0.032 (0.135)
	1	$\tilde{\beta}_2$	0.970 (0.030)	0.975 (0.030)	0.980 (0.023)	0.982 (0.023)
	1	$\tilde{\beta}_3$	0.978 (0.020)	0.981 (0.024)	0.984 (0.020)	0.986 (0.019)
Adaptive Lasso	1	$\tilde{\beta}_1$	0.060 (0.210)	0.041 (0.180)	0.044 (0.188)	0.019 (0.123)
	1	$\tilde{\beta}_2$	0.770 (0.350)	0.873 (0.280)	0.916 (0.023)	0.911 (0.243)
	1	$\tilde{\beta}_3$	0.810 (0.320)	0.900 (0.240)	0.928 (0.217)	0.926 (0.217)

Table 2 reported estimated metabolite concentrations ($\tilde{\boldsymbol{\beta}}$) corresponding to each level of shift variation across the four methods. Note that the proposed method defined $\tilde{\boldsymbol{\beta}}$ as the postselection error-corrected least squares estimates in (3.10). Additionally, $\tilde{\boldsymbol{\beta}}$ for each of the three regularized regression models: Lasso, elastic net, and adaptive Lasso, was defined as the estimates minimizing the corresponding loss function in Tibshirani (1996), Zou and Hastie (2005), and Zou (2006), respectively, where, $\tilde{\boldsymbol{\beta}}_{\text{Lasso}} = \arg\min_{\beta} \sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{p} \beta_j g_j(x_i; \mathbf{l}_j) \right\}^2 + \lambda \sum_{j=1}^{p} |\beta_j|$; $\tilde{\boldsymbol{\beta}}_{\text{elastic net}} = \arg\min_{\beta} \sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{p} \beta_j g_j(x_i; \mathbf{l}_j) \right\}^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$; and $\tilde{\boldsymbol{\beta}}_{\text{adaptive Lasso}} = \arg\min_{\beta} \sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{p} \beta_j g_j(x_i; \mathbf{l}_j) \right\}^2 + \lambda \sum_{j=1}^{p} \omega_j |\beta_j|$. As expected, the estimated coefficient for the shifted metabolite $M1$, i.e., $\tilde{\beta}_1$ was further away from its true concentration of 1 as the shifting errors increased. Particularly, $\tilde{\beta}_1$ decreased from 0.970 to 0.939; and from 0.017 to 0.002 for the proposed method and Lasso, respectively. Even at the smallest amount of shifting (±0.01 ppm), Lasso quantified the abundance of metabolite $M1$ with an estimate of 0.017. Compared to Lasso, elastic net and adaptive Lasso yielded slightly better estimates for β_1 of 0.111 and 0.06 respectively, yet still significantly underestimated the true parameter. On the other hand, the proposed method with an ability to detect and locate any potential peak shifting through the weight function in (2.5) provided a better estimate of 0.970.

Figure 2(b) depicts the weight plot for the $M1$ peak with the fixed search window of size $d = 10$, i.e., 0.01 ppm. In this illustration, the observed peak of $M1$ at 1.933 ppm is shifted from its referenced location at $M1_{\text{ref}}$ (i.e., 1.924 ppm). In principle, if a given peak does not shift, its corresponding weight plot would reach a maximum at the center while the weights for all neighboring locations diminish quickly. In contrast, for any locations, which were not peaks (e.g., points along the baseline), the weights would equally be distributed across all the points within the search window. As a result, such weight plots helped identify and relocate any shifted peaks so that they match their counterparts in the reference database as closely as possible. In the weight plot of Figure 2(b), the observed peak of $M1$ located at 1.933 ppm had the maximum weight (≈ 0.8) at 1.924 ppm. This suggests that the observed peak ($M1$) might have been generated from the reference $M1_{\text{ref}}$ and deviated from its referenced position by 0.009 ppm. Therefore, the reference spectrum of $M1_{\text{ref}}$ needs to be repositioned accordingly (Figure 2(c) and (d)) to ensure precise estimation.

5. Real data analysis

5.1. *Experimental mixtures*

Three experimental mixtures of different compositions of 20 amino acids, as outlined in Table S11 of the Supplementary material available at *Biostatistics* online, were used for performance comparisons across Lasso, elastic net, adaptive Lasso, Bayesil, Chenomx, and ASICS. The performances were evaluated using an increasing size of the reference library (61, 101, and 200, respectively) based on accuracy, sensitivity, and specificity. Based on the sensitivity analyses in Section S4 of the Supplementary material available at *Biostatistics* online, we set $d = 10$ (defined in (2.5)) and continued fixing $c_0 = 7\%$AUC (defined in Section 2.1) and $\sigma_0 = \max(y_i^2)/3$ (defined in (2.5)) for the real data analysis reported herein. Overall, the proposed method yielded the highest rate across the three metrics regardless of the library size.

As the number of candidate metabolites increased, it became easier to incorrectly claim the presence of metabolites. Table 3 showed a slight drop in specificity for both the proposed method (from 0.90 to 0.85) and Bayesil (from 0.66 to 0.13) in the mixture of all 20 amino acids. Interestingly, the automated profiler feature of Chenomx and ASICS often failed to capture some metabolites that were actually present in the mixture as shown by the number of false negatives (FN) in Table S12 of the Supplementary material available at *Biostatistics* online. As a result, both Chenomx and ASICS had the lowest sensitivity (0.80 and 0.45, respectively) compared to the remaining five methods (≈ 1) as shown in Table 3 for the mixture of 20 compounds. Unfortunately, we were not able to evaluate the impact of increasing library size from 101 to 200 on Bayesil since the maximum number of available metabolites was 93. Additionally, without the flexibility to adjust the reference library accompanying ASICS, which was fixed at 190, the impact of increasing library size on ASICS was not assessed.

Lasso, elastic net, and adaptive Lasso performed quite similarly in terms of producing relatively more false positives than the proposed method, which, in turn, led to a lower specificity. For example, in the last mixture of 20 compounds, the proposed method had a specificity of 0.85 while the three regularized models yielded the average rate of 0.73. For instance, L-lysine which was part of mixture 3 (Table S11 of the Supplementary material available at *Biostatistics* online), had one of its peaks located at 1.925 ppm. Even a slight shift of the peak in the observed spectrum could lead to a confusion with the one peak from acetic acid located at 1.924 ppm. Without considering shifting errors, it was not surprising to see Lasso, elastic net, and adaptive Lasso inaccurately classified acetic acid as being present in the mixture. Figure 3 (left) mirrored the observed and fitted spectra corresponding to the proposed method, Lasso, elastic net, and adaptive Lasso, respectively. The zoomed-in version in Figure S3 of the Supplementary material available at *Biostatistics* online depicted the false positive effects caused by the artifacts around 7.5 ppm on Lasso, elastic net, and adaptive Lasso.

The complexity level of a mixture spectrum also affected the accuracy and specificity. As the number of metabolites included in the mixture sample increased from 7 to 20 (Table 3), the accuracy decreased from 0.98 to 0.93 for the proposed method, from 0.93 to 0.81 for elastic net, and from 0.97 to 0.83 for Chenomx. Specificity encountered similar trends, with a drop from 0.97 to 0.85 for the proposed method, from 0.91 to 0.73 for the elastic net, and from 0.97 to 0.77 for Chenomx. Bayesil, Chenomx, and ASICS each used its own library of reference spectra which were collected at specific conditions (e.g., pH, temperature, etc.) to profile a given mixture spectrum. If the observed spectrum was collected under experimental conditions that were quite different from those in the reference libraries, it was possible

Table 3. *Comparison of proposed method with Lasso, elastic net, adaptive Lasso, using three experimental mixtures containing 6, 7, and 20 metabolites, respectively; and a library size of 61, 101, and 200 metabolites, respectively. Performance was evaluated based on average accuracy, sensitivity, and specificity*

# Met.	Metrics	Proposed method	Lasso	Elastic Net	Adaptive Lasso	Chenomx	Bayesil	ASICS
Library size 61								
6	Accuracy	1.00	0.72	0.62	0.75	0.80	0.64	0.80
	Sensitivity	1.00	1.00	1.00	1.00	0.67	1.00	0.57
	Specificity	1.00	0.69	0.58	0.73	0.82	0.60	0.81
7	Accuracy	1.00	0.77	0.79	0.64	0.89	0.64	0.83
	Sensitivity	1.00	1.00	1.00	1.00	0.71	1.00	0.33
	Specificity	1.00	0.74	0.76	0.59	0.91	0.59	0.86
20	Accuracy	0.93	0.67	0.69	0.70	0.74	0.75	0.56
	Sensitivity	1.00	1.00	1.00	1.00	0.80	0.95	0.45
	Specificity	0.90	0.51	0.54	0.56	0.71	0.66	0.59
Library size 101								
6	Accuracy	0.94	0.79	0.88	0.72	0.87	0.30	0.80
	Sensitivity	1.00	1.00	1.00	1.00	0.67	1.00	0.57
	Specificity	0.94	0.78	0.87	0.71	0.88	0.25	0.81
7	Accuracy	0.97	0.90	0.84	0.87	0.93	0.42	0.83
	Sensitivity	1.00	1.00	1.00	1.00	0.71	1.00	0.33
	Specificity	0.97	0.89	0.83	0.86	0.95	0.37	0.86
20	Accuracy	0.88	0.72	0.73	0.72	0.69	0.33	0.56
	Sensitivity	1.00	1.00	1.00	1.00	0.80	0.95	0.45
	Specificity	0.85	0.65	0.67	0.65	0.67	0.16	0.59
Library size 200								
6	Accuracy	0.94	0.86	0.86	0.83	0.92	0.30	0.80
	Sensitivity	1.00	1.00	1.00	1.00	0.67	1.00	0.57
	Specificity	0.93	0.85	0.85	0.82	0.93	0.25	0.81
7	Accuracy	0.98	0.86	0.86	0.92	0.96	0.42	0.83
	Sensitivity	1.00	1.00	1.00	1.00	0.71	1.00	0.33
	Specificity	0.97	0.85	0.85	0.92	0.97	0.37	0.86
20	Accuracy	0.86	0.74	0.75	0.74	0.76	0.33	0.56
	Sensitivity	1.00	1.00	1.00	1.00	0.80	0.95	0.45
	Specificity	0.84	0.71	0.72	0.71	0.75	0.16	0.59

that the induced shift variation was outside the range that these methods considered in their algorithms. This could consequentially cause failing to capture true metabolites (false negatives) and/or identifying wrong metabolites (false positives). Specifically, Table S13 of the Supplementary material available at *Biostatistics* online showed that Chenomx had a relatively lower number of false positives but it usually missed two to four metabolites across the three mixtures (e.g., L-alanine, L-cysteine, L-leucine, and L-glutamic acid). On the other hand, Bayesil only failed to identify at most one metabolite (i.e., L-aspartic acid in the third mixture) while having a larger number of false positives (around 14-22). Finally, ASICS

Table 4. *Comparison of proposed method with Lasso, Elastic net, adaptive Lasso, Chenomx, Bayesil, and ASICS using three serum samples from breast cancer patients (Hart and others, 2017) and a library size of 104 metabolites. Performance was evaluated based on average accuracy, sensitivity, and specificity over the three empirical samples (N = 3). Corresponding standard deviations are recorded in parentheses*

Metrics	Proposed method	Lasso	Elastic Net	Adaptive Lasso	Chenomx	Bayesil	ASICS
Accuracy	0.70 (0.01)	0.65 (0.03)	0.63 (0.04)	0.66 (0.03)	0.74 (0.02)	0.46 (0.03)	0.60 (0.01)
Sensitivity	0.67 (0.03)	0.67 (0.03)	0.68 (0.00)	0.67 (0.03)	0.05 (0.04)	0.55 (0.00)	0.54 (0.03)
Specificity	0.71 (0.01)	0.65 (0.04)	0.62 (0.05)	0.66 (0.05)	0.93 (0.04)	0.43 (0.04)	0.64 (0.02)

seemed to have both problems of missing true metabolites and identifying wrong metabolites. Figure 3 (right) mirrored the observed mixture spectrum with a fitted curve generated by the proposed method, Bayesil, Chenomx, and ASICS, respectively. More precisely, Figure S4 of the Supplementary material available at *Biostatistics* online zoomed in the spectral region from 1 to 4.5 ppm with the discrepancy between the observed and the profiled spectra to demonstrate the false positive and false negative effects on Bayesil, Chenomx, and ASICS as compared to the proposed approach.

5.2. *Biological samples*

Three serum samples of breast cancer patients from Hart *and others* (2017), which were publicly available on the MetaboLights database (Kale *and others*, 2016), were used to evaluate the practical application of the proposed method in comparison with the six other models. Evaluation criteria still included accuracy, sensitivity, and specificity, which were calculated using the metabolites manually identified by the authors of the study. According to Table 4, the proposed model yielded the best overall results with regard to identifying true metabolites in the mixtures while controlling the number of falsely identified metabolites. Similar to experimental results in Section 5.1, the six methods except for Chenomx identified more incorrect metabolites than the proposed approach. Specifically, the average sensitivity of our method was 0.70, while the remaining six approaches yielded an average sensitivity of less than 0.63. Across replications, Chenomx detected at most two metabolites out the total of 22 metabolites ($<10\%$) present in the mixture samples, which resulted in a low sensitivity of 0.05. Given the relatively low number of identified compounds, Chenomix, not surprisingly, produced the highest rate of specificity of 0.93.

Interestingly, there was a consistency across all methods with regard to failing to detect some of the metabolites that were previously identified by the authors of the original study. A signal-to-noise ratio (SNR) was calculated using TopSpin 4.06 (Bruker, Germany) for the mixture spectra. Three missed metabolites that included citric acid, formic acid, and phenylalanine, had a SNR less than 2. This is below the generally accepted lower-limits for peak detection.

ASICS had some built-in steps to exclude any reference metabolites from the library if at least one of its corresponding peaks did not show up in the observed complex spectrum. However, doing so could lead to eliminate some potential metabolites as it was quite possible for a given metabolite to lose one or more peaks due to experimental fluctuations (Zangger, 2015). This could potentially contributing to a large number of false negatives as seen in Table S13 of the Supplementary material available at *Biostatistics* online. Bayesil, on the other hand, did not integrate any regularization to reflect the sparsity of abundant metabolites in a complex mixture. As a result, Bayesil tended to identify more metabolites which were not part of the mixtures.

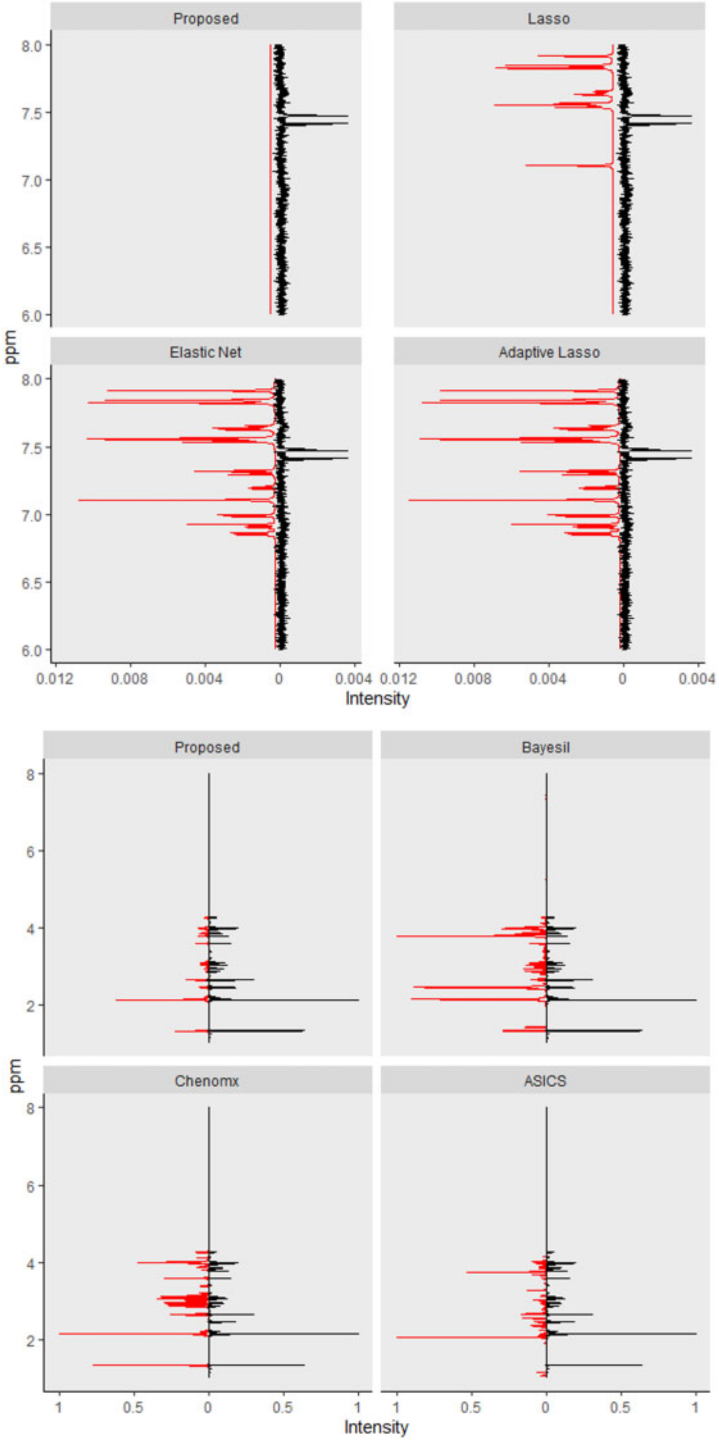

Fig. 3. *Top panel:* Each fitted curve (left) generated from the proposed method, Lasso, elastic net, and adaptive Lasso is mirrored with the observed mixture spectrum (right). *Bottom panel:* Each fitted curve (left) generated from the proposed method, Bayesil, Chenomx, and ASICS is mirrored with the observed mixture spectrum (right).

6. Conclusion

Metabolite identification and quantification play an essential role in any NMR metabolomics studies and are necessary to describe the underlying biological processes being investigated. However, manual assignment approaches are time-consuming, labor-intensive, and reliant on the knowledge and assessment of NMR experts. Many curve-fitting models with reference library support have been introduced to automate the metabolite identification process, but none has been unanimously demonstrated to be a gold standard approach in practice. We proposed a new approach that focused on addressing two major challenges of metabolite identification from complex mixtures: undesirable perturbations in signal locations and sparsity metabolites relative to reference databases. The proposed method was assessed using simulated, experimental and biological NMR metabolomics data sets. The overall performance was based on three metric including accuracy, sensitivity, and specificity. In addition, a comparison was made between the newly introduced approach and Lasso, elastic net, adaptive Lasso, Chenomx, Bayesil and ASICS with an increasing size of a reference library.

As a hybrid approach, the proposed leveraged sparsity properties from regularized models (e.g., Lasso, elastic net, and adaptive Lasso) which allowed a selection of potential metabolites from a large reference library. With the promising performance of our proposed method demonstrated using library size up to 200 (as shown in Table 3), this could potentially be scaled up to a relatively larger library. Additionally, our method incorporated a search window at each observed signal to capture any peak shifting which was then corrected for before estimating the metabolite concentration. Lastly, our approach with the modified objective function to correct for peak shifting and regularization enforcement was easy to implement. Though we empirically pre-selected some hyperparameters such as d, c_0, and σ_0, and σ_ϵ, we demonstrated in the sensitivity analyses (Tables S4–S7 of the Supplementary material available at *Biostatistics* online) that these values did not affect our method performance much.

The proposed method showed the best results in capturing metabolites that were truly present in the mixtures while keeping incorrect assignments at a relatively lower level regardless of increasing shifting variations as demonstrated in the simulation studies. In addition, as the complexity of a mixture spectrum increased and the number of candidate metabolites grew larger, all methods shared a common trend of tending to identify more incorrect metabolites. In other words, the combination of the two factors resulted in lower accuracy and specificity for all models. It was interesting to see that the automated Profiler feature of Chenomx only assigned a small number of metabolites to the mixtures. Such conservative assignments ensured fewer false identifications, which in turn led to a higher specificity overall. However, Chenomx often failed at capturing some true metabolites leading to a low sensitivity. On the contrary, Bayesil's aggressive detection yielded a large number of false positives while identifying most of the true positives. Consequently, Bayesil had a good sensitivity yet its specificity suffered considerably. Uniquely, ASICS seemed to encounter both problems of missing true metabolites and identifying wrong metabolites, resulting low sensitivity and specificity. Nonetheless, the new method still managed to maintain the best results across all three criteria (accuracy, sensitivity, and specificity) as compared to the others.

Even though we did not theoretically quantify the inferences of the regression coefficients in this article, we think it is possible to obtain proper confidence intervals for the coefficients, with special attention paid to the characteristics of the NMR spectral data. Again, as the final concentration estimation is conditional on both stages of the variable selection and covariate shifting correction, we utilize the random subsampling approach to closely investigate the stability of each stage. Based on the detailed results in Section 1.3 of the Supplementary material available at *Biostatistics* online, we observe that the two steps are stable in terms of selecting the true metabolites with a selection probability of at least 0.9, and estimating the covariate shifting error reasonably well. Such a stability procedure could be utilized to quantify the variability and construct confidence intervals for the metabolite concentration. Additionally, we could also leverage the jackknife resampling method, where each NMR peak is removed at a time to construct the jackknife confidence intervals. With regard to the nature of our proposed algorithm, the weights and the regression coefficients $\boldsymbol{\beta}$ are updated alternately at each coordinate descent step. More specifically, when the weights are updated using a new estimate of $\boldsymbol{\beta}$, the objective function may not decrease. As a result, the objective function does not always decrease monotonically after each step; and we have noticed this phenomenon in Figure S1 of the Supplementary material available at *Biostatistics* online. Although our proposed algorithm is also not guaranteed to converge, we find that the numerical convergence is achieved in our simulation studies and real data analyses. Additionally, we choose the initial values $\boldsymbol{\beta}^{(0)} = 0_{p\times 1}$ in our numerical analyses due to a practical reason that a majority of metabolites are not present in a mixture (i.e., $\beta_j = 0$). Alternatively, we could also experiment with multiple initializations and select the ones which yield the smallest objective value. The theoretical convergence properties of the proposed method is important but is beyond the scope of this article and will be investigated in our future work.

Software

Corresponding code of the method in the form of GNU Octave is incorporated in a toolkit called MVAPACK (Worley and Powers, 2014), which is publicly available for academic users at http://bionmr.unl.edu/mvapack.php. The equivalent R code is available at https://github.com/thaovu1/SCRR.

Supplementary material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors would like to thank the Associate Editor and two anonymous referees for their helpful and constructive comments, which led to a significant improvement of this article. We also thank Fatema Bhinderwala and Hannah Noel for generously providing us the experimental NMR mixture data sets.
Conflict of Interest: None declared.

Funding

The National Science Foundation (1660921); the Redox Biology Center (P30 GM103335, NIGMS), in part; and the Nebraska Center for Integrated Biomolecular Communication (P20 GM113126, NIGMS). The research was performed in facilities renovated with support from the National Institutes of Health (RR015468-01). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Ahmed, O. A. (2005) New denoising scheme for magnetic resonance spectroscopy signals. *IEEE Transactions on Medical Imaging*, **24**, 809–816.

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

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009) Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, **37**, 1705–1732.

Bühlmann, P. and Van de Geer, S. (2011) *Statistics for High-dimensional Data: Methods, Theory and Applications*. Berlin/Heidelberg, Germany: Springer Science & Business Media.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006) *Measurement Error in Nonlinear Models: A Modern Perspective*. Boca Raton, Florida, USA: Chapman and Hall/CRC.

Chenomx, N. (2015) *Suite.* Edmonton, AB, Canada: Chenomx Inc.

Datta, A. and Zou, H. (2017) Cocolasso for high-dimensional error-in-variables regression. *The Annals of Statistics*, **45**, 2400–2426.

Daviss, B. (2005) Growing pains for metabolomics: the newest'omic science is producing results–and more data than researchers know what to do with. *The Scientist*, **19**, 25–29.

Dieterle, F., Ross, A., Schlotterbeck, G. and Senn, H. (2006) Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. application in 1H NMR metabonomics. *Analytical Chemistry*, **78**, 4281–4290.

Dona, A. C., Kyriakides, M., Scott, F., Shephard, E. A., Varshavi, D., Veselkov, K. and Everett, J. R. (2016) A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. *Computational and Structural Biotechnology Journal*, **14**, 135–153.

Fiehn, O. (2002) Metabolomics—the link between genotypes and phenotypes. *Plant Mol Biol* **48**, 155–171.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007) Pathwise coordinate optimization. *The Annals of Applied Statistics*, **1**, 302–332.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1.

Gowda, G. N., Zhang, S., Gu, H., Asiago, V., Shanaiah, N. and Raftery, D. (2008) Metabolomics-based methods for early disease diagnostics. *Expert Review of Molecular Diagnostics*, **8**, 617–633.

Hart, C. D., Vignoli, A., Tenori, L., Uy, G. L., Van To, T., Adebamowo, C., Hossain, S. M., Biganzoli, L., Risi, E., Love, R. R. *and others* (2017) Serum metabolomic profiles identify ER-positive early breast cancer patients at increased risk of disease recurrence in a multicenter population. *Clinical Cancer Research*, **23**, 1422–1431.

Hollas, J. M. (2004) *Modern Spectroscopy*, 35–36. Hoboken, New Jersey, USA: Wiley.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) *An Introduction to Statistical Learning*, vol. 112. Berlin/Heidelberg, Germany: Springer.

Kale, N. S., Haug, K., Conesa, P., Jayseelan, K., Moreno, P., Rocca-Serra, P., Nainala, V. C., Spicer, R. A., Williams, M., Li, X. *and others* (2016) MetaboLights: an open-access database repository for metabolomics data. *Current Protocols in Bioinformatics*, **53**, 14–13.

Lefort, G., Liaubet, L., Canlet, C., Tardivel, P., Père, M.-C., Quesnel, H., Paris, A., Iannuccelli, N., Vialaneix, N. and Servien, R. (2019) ASICS: an R package for a whole analysis workflow of 1D 1H NMR spectra. *Bioinformatics*, **35**, 4356–4363.

Meinshausen, N. and Bühlmann, P. (2010) Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 417–473.

Psychogios, N., Hau, D.D., Peng, J., Guo, A.C., Mandal, R., Bouatra, S., Sinelnikov, I., Krishnamurthy, R., Eisner, R., Gautam, B. and Young, N. (2011) The human serum metabolome. *PLoS One*, **6**, e16957.

Putri, S. P., Nakayama, Y., Matsuda, F., Uchikata, T., Kobayashi, S., Matsubara, A. and Fukusaki, E. (2013) Current metabolomics: practical applications. *Journal of Bioscience and Bioengineering*, **115**, 579–589.

Ramirez, T., Daneshian, M., Kamp, H., Bois, F. Y., Clench, M. R., Coen, M., Donley, B., Fischer, S. M., Ekman, D. R., Fabian, E. *and others* (2013) Metabolomics in toxicology and preclinical research. *Altex*, **30**, 209.

RAVANBAKHSH, S., LIU, P., BJORDAHL, T. C., MANDAL, R., GRANT, J. R., WILSON, M., EISNER, R., SINELNIKOV, I., HU, X., LUCHINAT, C. AND OTHERS (2015) Accurate, fully-automated NMR spectral profiling for metabolomics. *PLoS One*, **10**, e0124219.

ROSENBAUM, M. AND TSYBAKOV, A. B. (2010) Sparse recovery under matrix uncertainty. *The Annals of Statistics*, **38**, 2620–2651.

SØRENSEN, Ø., FRIGESSI, A. AND THORESEN, M. (2015) Measurement error in LASSO: impact and likelihood bias correction. *Statistica Sinica*, **25**, 809–829.

SØRENSEN, Ø., HELLTON, K. H., FRIGESSI, A. AND THORESEN, M. (2018) Covariate selection in high-dimensional generalized linear models with measurement error. *Journal of Computational and Graphical Statistics*, **27**, 739–749.

TARDIVEL, P. J., CANLET, C., LEFORT, G., TREMBLAY-FRANCO, M., DEBRAUWER, L., CONCORDET, D. AND SERVIEN, R. (2017) ASICS: an automatic method for identification and quantification of metabolites in complex 1D 1 H NMR spectra. *Metabolomics*, **13**, 1–9.

THULIN, E., THULIN, M. AND ANDERSSON, D. I. (2017) Reversion of high-level mecillinam resistance to susceptibility in *Escherichia coli* during growth in urine. *EBioMedicine*, **23**, 111–118.

TIBSHIRANI, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 267–288.

TULPAN, D., LÉGER, S., BELLIVEAU, L., CULF, A. AND ČUPERLOVIĆ-CULF, M. (2011) MetaboHunter: an automatic approach for identification of metabolites from 1H-NMR spectra of complex mixtures. *BMC Bioinformatics*, **12**, 1–22.

VU, T., SIEMEK, P., BHINDERWALA, F., XU, Y. AND POWERS, R. (2019) Evaluation of multivariate classification models for analyzing NMR metabolomics data. *Journal of Proteome Research*, **18**, 3282–3294.

WELJIE, A. M., NEWTON, J., MERCIER, P., CARLSON, E. AND SLUPSKY, C. M. (2006) Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Analytical Chemistry*, **78**, 4430–4442.

WISHART, D. S. (2008a) Metabolomics: applications to food science and nutrition research. *Trends in Food Science & Technology*, **19**, 482–493.

— (2008b) Quantitative metabolomics using NMR. *TrAC Trends in Analytical Chemistry*, **27**, 228–237.

WISHART, D. S., FEUNANG, Y. D., MARCU, A., GUO, A. C., LIANG, K., VÁZQUEZ-FRESNO, R., SAJED, T., JOHNSON, D., LI, C., KARU, N. AND OTHERS. (2018) HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research*, **46**, D608–D617.

WORLEY, B. AND POWERS, R. (2014) MVAPACK: a complete data handling package for NMR metabolomics. *ACS Chemical Biology*, **9**, 1138–1144.

ZANGGER, K. (2015) Pure shift NMR. *Progress in Nuclear Magnetic Resonance Spectroscopy*, **86**, 1–20.

ZOU, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.

ZOU, H. AND HASTIE, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320.

[*Received April 19, 2021; revised April 11, 2022; accepted for publication April 24, 2022*]