Biostatistics (2022), **0**, 0, *pp.* S.1–S.21
doi:10.1093/biostatistics/output

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

THAO VU

Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA

YUHANG XU*

Department of Applied Statistics and Operations Research, Bowling Green State University, Bowling Green, OH 43402, USA

xuy@bgsu.edu

YUMOU QIU

Department of Statistics, Iowa State University, Ames, IA 50011, USA

ROBERT POWERS*

Department of Chemistry, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

rpowers3@unl.edu

*To whom correspondence should be addressed.

The Supplementary Material is organized as follows. Section S1 describes a detailed derivation of the objective function $L(\boldsymbol{\beta})$, the convergence criterion of the proposed method, and some properties of the post-selection estimates. Section S2 reports results of the additional simulation study in detail. Section S3 includes sensitivity analyses to assess the performance of the proposed method with different values of the tuning parameters such as search window size d, threshold level c_0, weight distributor σ_0, and noise standard deviation σ_ϵ. Finally, Section S4 contains supporting details for selecting d for experimental mixtures and illustrates the performance of each method with mirrored figures of the observed and reconstructed spectra.

S1. Additional details on the method implementation

S1.1 *Derivation*

Detailed derivation of finding derivatives of the first term $W(\beta)$ of the objective function $L(\boldsymbol{\beta})$ is as follows:

$$
\begin{aligned}
\frac{\partial}{\partial \beta_j} W(\beta) &= -\sum_{i=1}^{n}\sum_{k=k_l}^{k_u} w_{ik}\left\{y_i - \beta_0 - \sum_{j=1}^{p}\beta_j g_j(x_k;\mathbf{l}_j)\right\} g_j(x_k;\mathbf{l}_j) \\
&= -\left[\sum_{i=1}^{n}\sum_{k=k_l}^{k_u} w_{ik}\left\{y_i - \beta_0 - \sum_{h\neq j}^{p}\beta_h g_h(x_k;\mathbf{l}_h) - \beta_j g_j(x_k;\mathbf{l}_j)\right\} g_j(x_k;\mathbf{l}_j)\right] \\
&= -\left[\sum_{i=1}^{n}\sum_{k=k_l}^{k_u} w_{ik}\left\{y_i - \beta_0 - \sum_{h\neq j}^{p}\beta_h g_h(x_k;\mathbf{l}_h)\right\} g_j(x_k;\mathbf{l}_j) - \beta_j \sum_{i=1}^{n}\sum_{k=k_l}^{k_u} w_{ik}\{g_j(x_k;\mathbf{l}_j)\}^2\right] \\
&= -(\rho_j - \beta_j z_j)
\end{aligned}
$$

where $k_l = \min(1, i-d)$, $k_u = \max(i+d, n)$, and

$$
\begin{aligned}
\rho_j &= \sum_{i=1}^{n}\sum_{k=k_l}^{k_u} w_{ik}\left\{y_i - \beta_0 - \sum_{h\neq j}^{p}\beta_h g_h(x_k;\mathbf{l}_h)\right\} g_j(x_k;\mathbf{l}_j) \\
z_j &= \sum_{i=1}^{n}\sum_{k=k_l}^{k_u} w_{ik}\{g_j(x_k;\mathbf{l}_j)\}^2.
\end{aligned}
$$

Given a fixed w_{ik}, second derivative of $W(\beta)$ is as follows:

$$
\begin{aligned}
\frac{\partial^2}{\partial\beta_j\partial\beta_i}W(\beta) &= -\left(\frac{\partial}{\partial\beta_i}\rho_j - \beta_j\frac{\partial}{\partial\beta_i}z_j\right) \\
&= -\frac{\partial}{\partial\beta_i}\rho_j \\
&= \sum_{i=1}^{n}\sum_{k=k_l}^{k_u} w_{ik}g_j(x_k;\mathbf{l}_j)g_i(x_k;\mathbf{l}_i)
\end{aligned}
$$

with $g(), w_{ik} \geqslant 0$, $\frac{\partial^2}{\partial\beta_j\partial\beta_i}W(\beta) \geqslant 0$. Thus, at fixed w_{ik}, $W(\beta)$ is convex.

S1.2 *Convergence criterion*

As described in Section 3.1 of the main text, the process of the coordinate descent is repeated until the criterion $\|\hat{\boldsymbol{\beta}}^{(u)} - \hat{\boldsymbol{\beta}}^{(u-1)}\| < 10^{-5}$ is met or the maximum number of iterations, which is set at 1000 in our numerical analysis, is reached. Fig. S1 shows a decline of the loss function as the number of iterations increases using two representative simulated spectra (in Section 4), three experimental mixtures (in Section 5.1), and one representative biological serum (in Section 5.2). More specifically, the loss function is stabilized within the first 50 iterations across the mixture spectra, which verifies the convergence of the proposed method in practice.

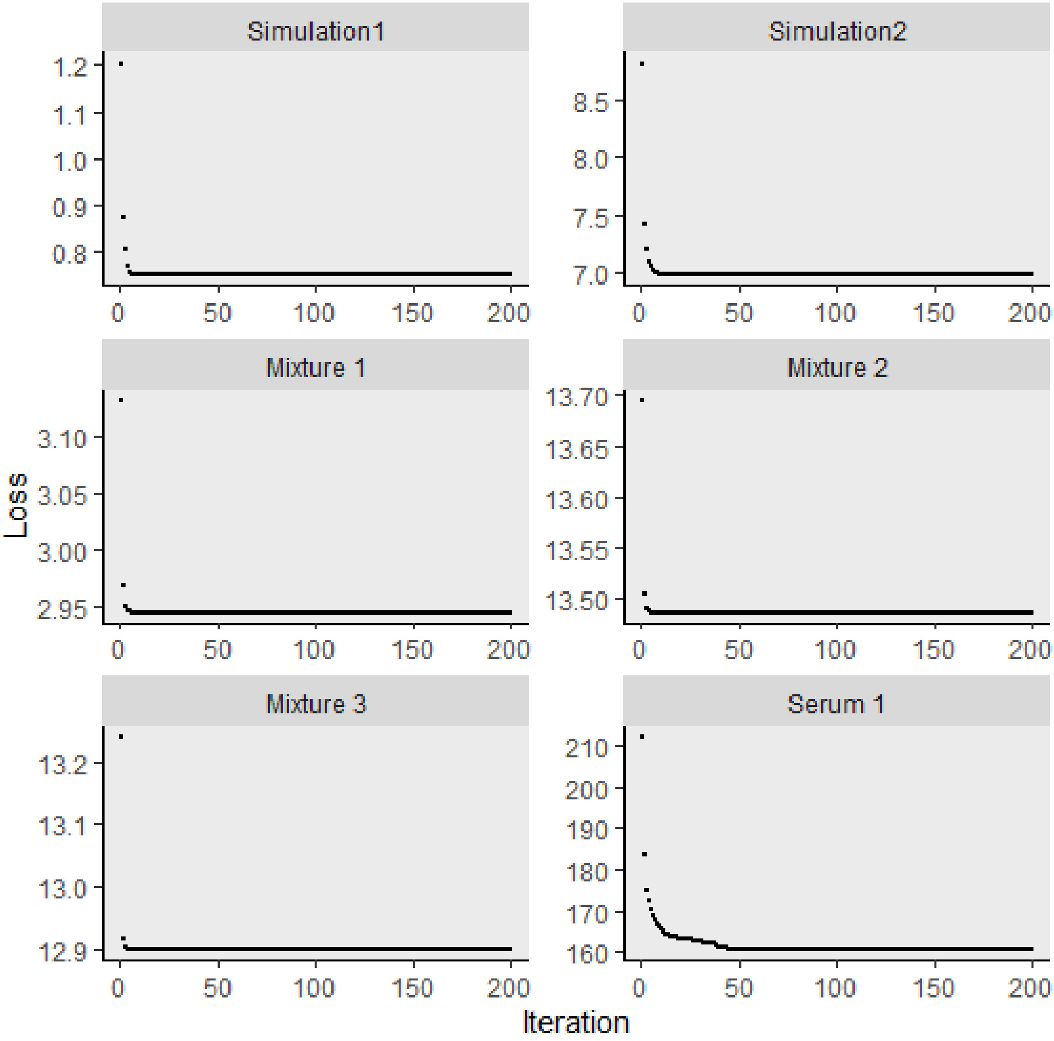

Fig. S1. Loss values are plotted as the number of iterations increases for two representative simulated spectra, three experimental mixtures, and one representative biological serum respectively.

S1.3 *Properties of post-selection least square estimates*

In order to evaluate the effect of the correction step on the post-selection least square estimator, we repeated the simulation in Section 4 for the proposed method without peak shift correction prior to the estimation procedure. In particular, similar to Section 3.3, with the active set of metabolites identified in the target mixture denoted as $\mathcal{A} = \{j : j \in \{1, \ldots, p\}, \hat{\beta}_j > 0\}$, the post-selection without-correction estimation was denoted as $\tilde{\boldsymbol{\beta}}^*$ 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

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

Table S1 reported the estimated metabolite concentration without shifting correction for the simulation at each level of shift variation using the proposed method. As expected, the bias of the estimated concentration for the shifted metabolite (i.e., $\tilde{\beta}^*_1$) was much larger than that of the remaining two metabolites. In contrast, by imposing the shifting correction step prior to quantification, the bias of the post-selection $\tilde{\beta}_1$ (Table 2) was much smaller than $\tilde{\beta}^*_1$. In other words, the bias in the estimation obtained from the proposed method is due to the nature of the quantification procedure.

The least square estimate at the quantification step is both post-selected and post-corrected. It relies on the variable selection result and the correction for shifting error. This is more complicated than the usual post-selection inference of the Lasso estimators. To understand the impact of the two steps on the least square estimator, we could consider the stability selection procedure, as in Meinshausen and Bühlmann (2010). For the simulation reported in Section 4, setting the shift of the first metabolite fixed at 0.005 ppm, i.e., $\delta_{11} = 0.005$, we created $N = 300$ subsamples, each with size of $n^* = \lfloor n/2 \rfloor$, by drawing a random subset of $\{1, \ldots, n\}$ without replacement, as n represented the total of all data points. More specifically, for each ith subsample, $i = 1, \ldots, N$, the process of cross validating to select an optimal regulation parameter $\lambda^{(i)}_{opt}$ and fitting the model to obtain the first-stage estimated coefficients $\hat{\boldsymbol{\beta}}^{(i)}$ was performed. A selection set for the ith subsample was denoted as $S^{(i)} = \{j : \hat{\beta}^{(i)}_k > 0, j \in \{1, 2, \ldots, 200\}\}$. The empirical selection probability for each jth model component was calculated as the average number of times the jth metabolite was selected, i.e., $\pi_j = \sum_{i=1}^{N} I(j \in S^{(i)})/N$, where $I(.)$ denoted the indicator function. Finally, the stable set was obtained such that $\hat{S}^{stable} = \{j : \pi_j > \pi_{thr}\}$ with π_{thr}

as a threshold. We obtained $\hat{S}^{stable} = \{1, 2, 3\}$, consisted metabolite indexes in the set of 200 metabolites with the corresponding selection probability above the threshold level $\pi_{thr} = 0.9$. Recall the true parameters in the simulation were set up such that such that $\beta_1 = \beta_2 = \beta_3 = 1$, and $\beta_j = 0$ for $j = 4, \ldots, 200$. This result showed the stability of our proposed method in selecting the true metabolites. We continued the second stage involving covariate shifting correction for each subsample. In particular, for each ith subsample, we estimated the shift for the mth peak of the jth metabolite as in Section 3.3 of the paper, denoted as $\hat{\delta}_{jm}^{(i)}$, for $m = 1, \ldots, n_j$ with n_j as the total number of peaks for jth metabolite. The estimated shift was then averaged across N subsamples, i.e., $\hat{\delta}_{jm}^{stable} = (1/N) \sum_{i=1}^{N} \hat{\delta}_{jm}^{(i)}$. For the first metabolite in the selection set, we obtained $\hat{\delta}_{11}^{stable} = 0.004$ (with standard deviation of 0.001), which was close to the true shift $\delta_{11} = 0.005$. Additionally, the method estimated the shift for the second and third metabolites, which were not actually shifted, to be close to 0 (with the average shifts of 0.0001 and 0.0005, respectively).

Table S 1. Average estimated metabolite concentrations for 200 iterations in the simulation described in Section 4 at an increasing shifting variations from ±0.01 ppm to ±0.04 ppm using the proposed method without shift correction prior to the quantification step. Corresponding standard deviations are recorded in parentheses

	Truth	$\tilde{\boldsymbol{\beta}}^*$	± 0.01ppm	±0.02ppm	±0.03ppm	±0.04ppm
Proposed Method	1	$\tilde{\beta}_1^*$	0.112 (0.230)	0.061 (0.185)	0.049 (0.152)	0.041 (0.157)
	1	$\tilde{\beta}_2^*$	0.995 (0.071)	0.999 (0.009)	0.999 (0.008)	1.000 (0.009)
	1	$\tilde{\beta}_3^*$	0.996 (0.071)	0.999 (0.006)	1.000 (0.006)	1.000 (0.006)

S2. Additional simulation study

The mixture complexity in the additional simulation study was increased by incorporating chemical compounds with multiplets, which were multiple resonances attributed to a single signal (as in Fig. S2 (a)). Specifically, the sample mixture spectrum was created by combining five individual spectra with random systematic variation. Specifically, $\beta_j = 1$ for $j = 1, \ldots, 5$, and $\beta_j = 0$

for $j = 6, \ldots, 200$. As in Fig. S2 (a), the mixture spectrum was shown in black with multiple clusters of signals. The overlaid red curve represented the reference spectrum of compound 2. We intentionally varied the first cluster of two peaks (i.e., doublet) together by an increasing amount $\{\delta_{2m}\}_{m=1}^{2} \sim \text{Unif}(-K, K)$, such that $K = \{0.01, 0.02, 0.03, 0.04\}$ ppm respectively to assess the performance of various methods. Overall, the proposed method outperformed Lasso, elastic net, and adaptive Lasso in identification accuracy, sensitivity and specificity. Additionally, by leveraging the weight matrix to relocate any potential peak shifting (as shown in Fig. S2 (b), (c)), the proposed method produced better estimated concentration for the shifted metabolite, $\hat{\beta}_2$.

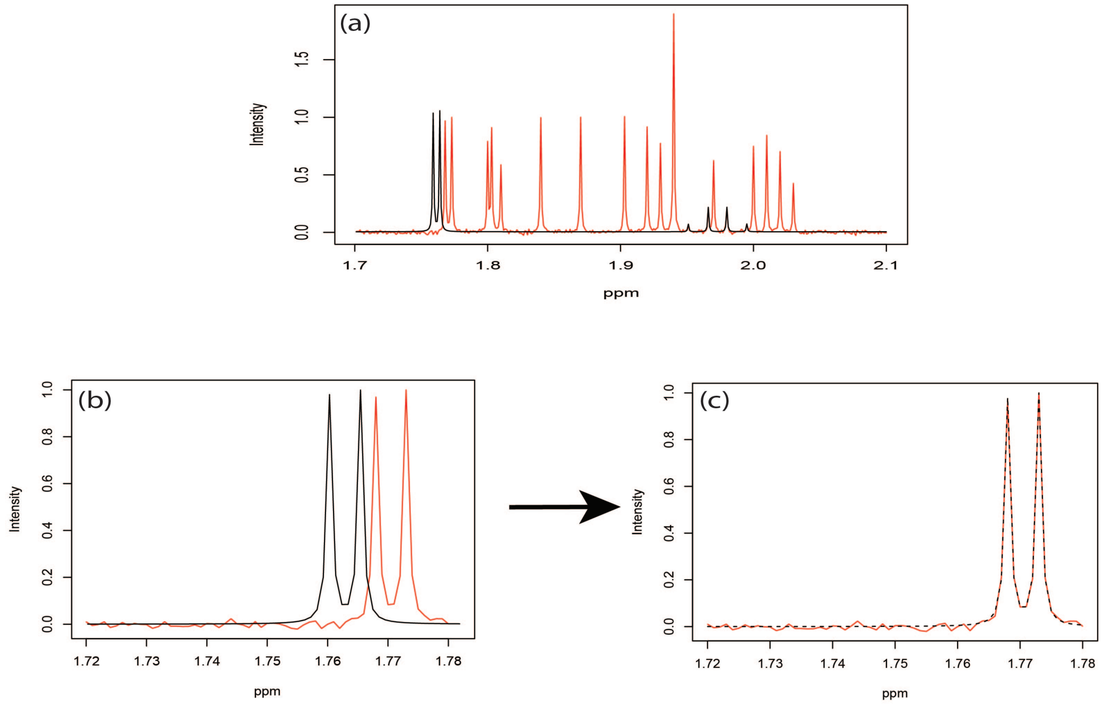

Fig. S2. Simulated mixture spectrum (black) with added random noise (a) overlaid with reference spectrum of the compound 2 (red) in the additional simulation. Shifted doublet (b) gets relocated as illustrated in (c) utilizing the weight matrix.

Table S2 reported results for the four methods across all levels of shifting errors. Accuracy, sensitivity, and specificity were decreasing functions of the positional variations. As peaks were shifted by a larger amount, it became harder for any model to correctly detect the presence of the corresponding metabolite in the mixture. Again, the proposed method produced the highest values among all metrics, while the elastic net consistently had the lowest accuracy regardless of the amount of shifting. Interestingly, sensitivity dropped faster across all four methods as shifting variations increased from ±0.01 to ±0.04 ppm; 1%, 9.8%,10%, and 12% for the proposed method, Lasso, elastic net, and adaptive Lasso, respectively.

Table S 2. Average accuracy, sensitivity, and specificity for 200 iterations in the additional simulation study 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 (0.001)	0.999 (0.002)	0.999 (0.002)	0.998 (0.003)
	Sensitivity	0.999 (0.001)	0.999 (0.001)	0.999 (0.014)	0.988 (0.052)
	Specificity	0.999 (0.001)	0.999 (0.002)	0.999 (0.002)	0.999 (0.002)
Lasso	Accuracy	0.991 (0.014)	0.991 (0.011)	0.989 (0.013)	0.988 (0.014)
	Sensitivity	0.943 (0.201)	0.936 (0.144)	0.890 (0.148)	0.850 (0.154)
	Specificity	0.993 (0.014)	0.992 (0.011)	0.991 (0.013)	0.991 (0.014)
Elastic net	Accuracy	0.967 (0.023)	0.970 (0.023)	0.966 (0.023)	0.969 (0.022)
	Sensitivity	0.998 (0.020)	0.955 (0.084)	0.916 (0.099)	0.889 (0.100)
	Specificity	0.966 (0.023)	0.970 (0.024)	0.967 (0.023)	0.971 (0.022)
Adaptive Lasso	Accuracy	0.992 (0.012)	0.990 (0.013)	0.989 (0.012)	0.989 (0.010)
	Sensitivity	0.993 (0.037)	0.944 (0.090)	0.900 (0.100)	0.874 (0.097)
	Specificity	0.992 (0.013)	0.992 (0.013)	0.990 (0.012)	0.991 (0.011)

Table S 3. Average estimated metabolite concentrations for 200 iterations in the additional 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	$\tilde{\beta}_1$	0.985 (0.019)	0.988 (0.017)	0.983 (0.030)	0.975 (0.053)
	1	$\tilde{\beta}_2$	0.838 (0.234)	0.920 (0.185)	0.751 (0.328)	0.633 (0.374)
	1	$\tilde{\beta}_3$	0.982 (0.023)	0.985 (0.022)	0.968 (0.088)	0.838 (0.281)
	1	$\tilde{\beta}_4$	0.895 (0.143)	0.940 (0.120)	0.872 (0.147)	0.817 (0.171)
	1	$\tilde{\beta}_5$	0.952 (0.049)	0.948 (0.044)	0.900 (0.086)	0.858 (0.106)
Lasso	1	$\tilde{\beta}_1$	0.930 (0.197)	0.962 (0.122)	0.968 (0.121)	0.989 (0.162)
	1	$\tilde{\beta}_2$	0.134 (0.133)	0.084 (0.132)	0.051 (0.106)	0.043 (0.104)
	1	$\tilde{\beta}_3$	0.916 (0.202)	0.951 (0.124)	0.960 (0.121)	0.952 (0.139)
	1	$\tilde{\beta}_4$	0.941 (0.200)	0.974 (0.125)	0.976 (0.122)	0.968 (0.140)
	1	$\tilde{\beta}_5$	0.938 (0.196)	0.972 (0.122)	0.979 (0.122)	0.975 (0.141)
Elastic net	1	$\tilde{\beta}_1$	0.985 (0.008)	0.986 (0.007)	0.987 (0.008)	1.010 (0.068)
	1	$\tilde{\beta}_2$	0.215 (0.223)	0.104 (0.200)	0.083 (0.199)	0.050 (0.144)
	1	$\tilde{\beta}_3$	0.980 (0.010)	0.981 (0.009)	0.981 (0.009)	0.981 (0.011)
	1	$\tilde{\beta}_4$	0.996 (0.016)	0.998 (0.019)	0.996 (0.018)	0.994 (0.016)
	1	$\tilde{\beta}_5$	0.990 (0.007)	0.994 (0.010)	0.996 (0.010)	0.998 (0.013)
Adaptive Lasso	1	$\tilde{\beta}_1$	0.972 (0.032)	0.981 (0.021)	0.981 (0.024)	1.002 (0.070)
	1	$\tilde{\beta}_2$	0.196 (0.221)	0.096 (0.198)	0.075 (0.199)	0.144 (0.140)
	1	$\tilde{\beta}_3$	0.960 (0.046)	0.972 (0.030)	0.971 (0.034)	0.975 (0.019)
	1	$\tilde{\beta}_4$	0.983 (0.032)	0.993 (0.027)	0.990 (0.029)	0.991 (0.019)
	1	$\tilde{\beta}_5$	0.981 (0.025)	0.990 (0.019)	0.992 (0.021)	0.997 (0.016)

The average estimated coefficients of compound concentrations for 200 runs and corresponding standard deviations were recorded in Table S3 for the four methods. Taking advantage of the weight function to relocate shifted peaks, our proposed method obtained a better estimation of concentration ($\tilde{\beta}_2$) for compound 2 compared to other approaches. Specifically, it estimated β_2 to be between 0.63 and 0.92 while Lasso had its estimation between 0.043 and 0.134. Compared to Lasso, elastic net and adaptive Lasso performed slightly better in terms of estimation (between 0.05 and 0.215; 0.144 and 0.196 respectively), though they were still far below the truth.

S3. Sensitivity Analysis

S3.1 *Assessment of different values of upper bound d simultaneously with threshold level c_0 and weight distributor σ_0*

According to Section 2.3 of the manuscript, d served as an upper bound index capturing the shifting errors $\{\delta_{jm}\}_{m=1}^{n_j}$ with $j = 1, \dots, p$, in which each $\{\delta_{jm}\}$ was bounded on $[-K, K]$ for a positive constant K ppm. This ensured the locality of shifting errors associated with signals in the mixture spectrum. With equal space of 0.001 ppm between chemical shifts, we can translate amount of shift K to d such that $\frac{K}{0.001} \leqslant d$.

As d played a critical role in our proposed approach, we designed sensitivity analyses to assess how d affected the method performance in conjunction with each of the remaining parameters including c_0 and σ_0. More precisely, we extended the first case of the simulation described in the main text with shifting variations within ± 0.01ppm to assess the performance of the proposed method with $d = 5$, $d = 10$, $d = 15$, $d = 20$, and $d = 25$. Tables S4–S7 herein reported results regarding accuracy, sensitivity, specificity, and estimated metabolite concentration for each combination set. Optimal results were observed for $d = 10$, $d = 15$, and $d = 20$. In particular, given a fixed value of c_0 or σ_0, accuracy, sensitivity, and specificity remained consistent and close to the optimal value of 1 for the three values of d. Moreover, estimated coefficients $\tilde{\beta}_1$, $\tilde{\beta}_2$, and $\tilde{\beta}_3$ of the three metabolites truly belonged to the mixture also stayed close to the true value of 1 across the three values of d. Therefore, in each simulation scenario corresponding to each range of shifting variation of $K = \{0.01, 0.02, 0.03, 0.04\}$ ppm, we fixed d to be the closest integer which captured the magnitude of maximum variation to reduce computational time. In other words, d was set to be 10, 20, 30, and 40, respectively.

As mentioned in Section 2.1, low-intensity NMR signals were not reliable for identifying metabolites in a complex mixture since they were likely to be generated from undesirable experimental perturbations. Therefore, for a given target spectrum, we only retained all signals above 7% of area under the spectrum (AUC), i.e., $c_0 = 7\% \times \text{AUC}$. Tables S4 and S5 recorded the performance of the proposed method in terms of accuracy, sensitivity, specificity, and estimated metabolite concentration using different c_0 values (i.e., 5%, 7%, 10%, and 12% of AUC) in conjunction with various d values (i.e., 5, 10, 15, 20, and 25). Generally, given a fixed d value, sensitivity and specificity were mostly unchanged while sensitivity slightly decreased as the threshold level c_0 increased (Table S4). Similar patterns were observed for the estimated concentration of true metabolites $\tilde{\beta}_1$, $\tilde{\beta}_2$, and $\tilde{\beta}_3$ (Table S5). This served as an assurance to continue our simulation studies as well as real data analysis using $c_0 = 7\%\text{AUC}$.

In a similar manner, we also examined how different values of σ_0 (i.e., $\max(y_i^2)/3$, $\max(y_i^2)/4$, $\max(y_i^2)/5$, and $\max(y_i^2)/6$) and various d values (i.e., 5, 10, 15, 20, and 25) would affect the method performance (Tables S6- S7). Again, optimal results were observed for $d = 10$, $d = 15$, and $d = 20$. In particular, the evaluation metrics including accuracy, sensitivity, and specificity remained consistently close to 1 as σ_0 decreased from $\max(y_i^2)/3$ to $\max(y_i^2)/6$. Moreover, the estimated metabolite concentration $\tilde{\beta}_1$, $\tilde{\beta}_2$, and $\tilde{\beta}_3$ also stayed close to the true value of 1. This again indicated that it was reasonable to carry on our analysis with the predefined $\sigma_0 = \max(y_i^2)/3$.

Table S 4. Average accuracy, sensitivity, and specificity for 200 iterations in the sensitivity analysis of different threshold values c_0 in conjunction with different d value, using shifting variation of ±0.01 ppm. Shaded row shows the results reported in the main text. Corresponding standard deviations are recorded in parentheses

	Metrics	$d = 5$	$d = 10$	$d = 15$	$d = 20$	$d = 25$
$c_0 = 5\%\text{AUC}$	Accuracy	0.996 (0.005)	0.998 (0.003)	0.998 (0.003)	0.996 (0.005)	0.995 (0.006)
	Sensitivity	0.752 (0.366)	0.998 (0.024)	0.998 (0.024)	0.998 (0.024)	0.965 (0.155)
	Specificity	0.999 (0.001)	0.999 (0.002)	0.999 (0.002)	0.997 (0.004)	0.996 (0.005)
$c_0 = 7\%\text{AUC}$	Accuracy	0.997 (0.005)	0.999 (0.002)	0.998 (0.003)	0.997 (0.004)	0.996 (0.005)
	Sensitivity	0.837 (0.304)	0.998 (0.024)	0.998 (0.024)	0.998 (0.024)	0.937 (0.193)
	Specificity	0.999 (0.001)	0.999 (0.002)	0.999 (0.003)	0.997 (0.004)	0.997 (0.004)
$c_0 = 10\%\text{AUC}$	Accuracy	0.996 (0.006)	0.999 (0.002)	0.999 (0.003)	0.996 (0.004)	0.995 (0.006)
	Sensitivity	0.763 (0.370)	0.999 (0.001)	0.998 (0.024)	0.999 (0.001)	0.953 (0.164)
	Specificity	0.999 (0.001)	0.999 (0.002)	0.999 (0.002)	0.996 (0.004)	0.996 (0.005)
$c_0 = 12\%\text{AUC}$	Accuracy	0.997 (0.005)	0.999 (0.002)	0.998 (0.003)	0.996 (0.004)	0.995 (0.005)
	Sensitivity	0.822 (0.317)	0.999 (0.001)	0.999 (0.001)	0.999 (0.001)	0.923 (0.223)
	Specificity	0.999 (0.001)	0.999 (0.002)	0.998 (0.003)	0.997 (0.004)	0.997 (0.004)

Table S 5. Average estimated metabolite concentrations for 200 iterations in the sensitivity analysis of different threshold values c_0 in conjunction with different d value, using shifting variation of ±0.01 ppm. Shaded row shows the results reported in the main text. Corresponding standard deviations are recorded in parentheses

	$\tilde{\boldsymbol{\beta}}$	$d = 5$	$d = 10$	$d = 15$	$d = 20$	$d = 25$
$c_0 = 5\%\text{AUC}$	$\tilde{\beta}_1$	0.476 (0.471)	0.997 (0.003)	0.997 (0.003)	0.997 (0.003)	0.823 (0.379)
	$\tilde{\beta}_2$	0.703 (0.419)	0.987 (0.071)	0.987 (0.071)	0.988 (0.071)	0.965 (0.156)
	$\tilde{\beta}_3$	0.695 (0.434)	0.994 (0.006)	0.994 (0.006)	0.995 (0.006)	0.973 (0.140)
$c_0 = 7\%\text{AUC}$	$\tilde{\beta}_1$	0.589 (0.478)	0.997 (0.003)	0.997 (0.003)	0.997 (0.003)	0.754 (0.429)
	$\tilde{\beta}_2$	0.831 (0.341)	0.987 (0.071)	0.987 (0.071)	0.987 (0.071)	0.947 (0.195)
	$\tilde{\beta}_3$	0.826 (0.355)	0.994 (0.006)	0.994 (0.006)	0.994 (0.006)	0.957 (0.176)
$c_0 = 10\%\text{AUC}$	$\tilde{\beta}_1$	0.511(0.481)	0.997 (0.003)	0.997 (0.003)	0.998 (0.003)	0.789 (0.406)
	$\tilde{\beta}_2$	0.746 (0.406)	0.991 (0.010)	0.986 (0.071)	0.992 (0.010)	0.965 (0.148)
	$\tilde{\beta}_3$	0.736 (0.415)	0.994 (0.006)	0.995 (0.006)	0.995 (0.006)	0.973 (0.143)
$c_0 = 12\%\text{AUC}$	$\tilde{\beta}_1$	0.479 (0.491)	0.997 (0.003)	0.998 (0.003)	0.998 (0.003)	0.739 (0.438)
	$\tilde{\beta}_2$	0.847 (0.335)	0.992 (0.009)	0.986 (0.009)	0.992 (0.009)	0.936 (0.214)
	$\tilde{\beta}_3$	0.834 (0.353)	0.995 (0.006)	0.995 (0.006)	0.995 (0.006)	0.943 (0.210)

Table S 6. Average accuracy, sensitivity, and specificity for 200 iterations in the sensitivity analysis of different values of σ_0 in conjunction with various d values, using shifting variation of ±0.01 ppm. Shaded row shows the results reported in the main text. Corresponding standard deviations are recorded in parentheses

	Metrics	$d = 5$	$d = 10$	$d = 15$	$d = 20$	$d = 25$
$\sigma_0 = \frac{\max(y_i^2)}{3}$	Accuracy	0.997 (0.005)	0.998 (0.003)	0.998 (0.003)	0.996 (0.005)	0.996 (0.006)
	Sensitivity	0.803 (0.329)	0.999 (0.081)	0.999 (0.001)	0.999 (0.001)	0.950 (0.176)
	Specificity	0.999 (0.001)	0.999 (0.003)	0.998 (0.003)	0.996 (0.005)	0.997 (0.004)
$\sigma_0 = \frac{\max(y_i^2)}{4}$	Accuracy	0.997 (0.005)	0.999 (0.003)	0.998 (0.004)	0.996 (0.004)	0.996 (0.006)
	Sensitivity	0.815 (0.315)	0.997 (0.047)	0.999 (0.001)	0.999 (0.001)	0.999 (0.001)
	Specificity	0.999 (0.001)	0.999 (0.003)	0.998 (0.003)	0.996 (0.004)	0.996 (0.004)
$\sigma_0 = \frac{\max(y_i^2)}{5}$	Accuracy	0.998 (0.004)	0.998 (0.003)	0.998 (0.004)	0.996 (0.004)	0.996 (0.006)
	Sensitivity	0.877 (0.235)	0.997 (0.047)	0.999 (0.001)	0.999 (0.001)	0.999 (0.001)
	Specificity	0.999 (0.001)	0.999 (0.003)	0.998 (0.003	0.996 (0.004)	0.996 (0.004)
$\sigma_0 = \frac{\max(y_i^2)}{6}$	Accuracy	0.998 (0.003)	0.998 (0.003)	0.998 (0.003)	0.996 (0.005)	0.996 (0.006)
	Sensitivity	0.905 (0.181)	0.997 (0.047)	0.999 (0.001)	0.980 (0.128)	0.999 (0.001)
	Specificity	0.999 (0.001)	0.999 (0.003)	0.998 (0.003)	0.997 (0.004)	0.996 (0.004)

Table S 7. Average estimated metabolite concentrations for 200 iterations in the sensitivity analysis of different values of σ_0 in conjunction with various d values, using shifting variation of ±0.01 ppm. Shaded row shows the results reported in the main text. Corresponding standard deviations are recorded in parentheses

	$\tilde{\boldsymbol{\beta}}$	$d = 5$	$d = 10$	$d = 15$	$d = 20$	$d = 25$
$\sigma_0 = \frac{\max(y_i^2)}{3}$	$\tilde{\beta}_1$	0.531 (0.481)	0.998 (0.003)	0.998 (0.003)	0.998 (0.003)	0.794 (0.403)
	$\tilde{\beta}_2$	0.785 (0.372)	0.977 (0.121)	0.993 (0.009)	0.992 (0.010)	0.960 (0.166)
	$\tilde{\beta}_3$	0.754 (0.394)	0.980 (0.121)	0.995 (0.006)	0.996 (0.006)	0.964 (0.160)
$\sigma_0 = \frac{\max(y_i^2)}{4}$	$\tilde{\beta}_1$	0.534 (0.484)	0.997 (0.003)	0.997 (0.003)	0.997 (0.003)	0.988 (0.099)
	$\tilde{\beta}_2$	0.813 (0.356)	0.987 (0.071)	0.992 (0.010)	0.991 (0.009)	0.993 (0.010)
	$\tilde{\beta}_3$	0.787 (0.372)	0.990 (0.071)	0.995 (0.007)	0.995 (0.006)	0.995 (0.006)
$\sigma_0 = \frac{\max(y_i^2)}{5}$	$\tilde{\beta}_1$	0.538 (0.492)	0.997 (0.003)	0.997(0.003)	0.997 (0.003)	0.998 (0.003)
	$\tilde{\beta}_2$	0.893 (0.272)	0.987 (0.071)	0.992 (0.010)	0.991 (0.009)	0.993 (0.010)
	$\tilde{\beta}_3$	0.894 (0.267)	0.990 (0.071)	0.995 (0.007)	0.995 (0.006)	0.995 (0.006)
$\sigma_0 = \frac{\max(y_i^2)}{6}$	$\tilde{\beta}_1$	0.543 (0.495)	0.997 (0.003)	0.997 (0.003)	0.987 (0.100)	0.998 (0.003)
	$\tilde{\beta}_2$	0.930 (0.209)	0.987 (0.071)	0.991 (0.009)	0.968 (0.156)	0.993 (0.010)
	$\tilde{\beta}_3$	0.949 (0.165)	0.990 (0.071)	0.994 (0.006)	0.970 (0.156)	0.995 (0.006)

S3.2 *Performance of the proposed method with an increasing noise level σ_ϵ*

Signal-to-noise ratio (SNR), shown in Table S8, was calculated for both simulations studies at an increasing shifting variation from ±0.01 ppm to ±0.04 ppm for 200 iterations, according to the formula introduced in Czanner *and others* (2008), while keeping the standard deviation of the errors ϵ_i at 5% of the area under the mixture spectrum curve (AUC), i.e., $\sigma_\epsilon = 5\%\text{AUC}$.

Table S 8. Average SNR for 200 iterations in both simulation studies at an increasing shifting variation from ±0.01 ppm to ±0.04 ppm. Corresponding standard deviations are recorded in parentheses

	±0.01 ppm	±0.02 ppm	±0.03 ppm	±0.04 ppm
Simulation	16.403 (4.628)	17.065 (3.718)	17.221 (3.740)	17.170 (3.917)
Additional simulation	6.584 (4.604)	5.064 (3.336)	5.095(3.344)	3.873(3.146)

Additionally, to further assess how different noise levels would affect the performance of the proposed method, we extended the simulation reported in the main text, using an increasing noise level, by multiplying the original noise standard deviation σ_ϵ with 1.2, 1.5, 1.7, and 2, respectively. Tables S9 and S10 reported the accuracy, sensitivity, specificity, and estimated metabolite concentration for the proposed method at different shifting variations and error standard deviations. Generally, the evaluation metrics including accuracy, sensitivity, and specificity decreased as the noise errors increased. At small shifting variations such as ±0.01 ppm and ±0.02 ppm, accuracy and specificity stayed consistently close to 1 while sensitivity slightly dropped. At shifting variations of ±0.03 ppm and ±0.04 ppm, accuracy and specificity experienced a small drop whereas sensitivity decreased by a relatively larger amount as the spectral noise increased. This was expected since the signal-to-noise ratio decreased with the increasing level of spectral noise. As a result, more artifact signals were considered in the model as real signals, leading to more false positives, thus, lower sensitivity. Similarly, estimated metabolite concentration $\tilde{\beta}_1$, $\tilde{\beta}_2$, $\tilde{\beta}_3$ also decreased as the spectral noise increased across all four levels of shifting variations. It was not surprising to observe a relatively bigger drop in the $\tilde{\beta}$'s at the combination of highest level of spectral noise and shifting variation (toward the bottom right corner of Table S10).

Table S 9. Average accuracy, sensitivity, and specificity for 200 iterations in the sensitivity analysis of different values of σ_0 in conjunction with various d values, using shifting variation of ±0.01 ppm. Shaded row shows the results reported in the main text. Corresponding standard deviations are recorded in parentheses

	Metrics	±0.01 ppm	±0.02 ppm	±0.03 ppm	±0.04 ppm
$1.0\sigma_\epsilon$	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)
$1.2\sigma_\epsilon$	Accuracy	0.999 (0.002)	0.998 (0.003)	0.997 (0.005)	0.995 (0.006)
	Sensitivity	0.993 (0.066)	0.992 (0.062)	0.978 (0.134)	0.973 (0.154)
	Specificity	0.999 (0.002)	0.998 (0.003)	0.997 (0.004)	0.996 (0.006)
$1.5\sigma_\epsilon$	Accuracy	0.999 (0.002)	0.998 (0.003)	0.995 (0.006)	0.994 (0.007)
	Sensitivity	0.997 (0.047)	0.998 (0.024)	0.968 (0.155)	0.938 (0.232)
	Specificity	0.999 (0.002)	0.998 (0.003)	0.996 (0.006)	0.994 (0.007)
$1.7\sigma_\epsilon$	Accuracy	0.999 (0.002)	0.997 (0.004)	0.995 (0.006)	0.993 (0.007)
	Sensitivity	0.995 (0.071)	0.992 (0.062)	0.930 (0.226)	0.930 (0.245)
	Specificity	0.999 (0.002)	0.997 (0.004)	0.996 (0.005)	0.994 (0.007)
$2.0\sigma_\epsilon$	Accuracy	0.999 (0.002)	0.997 (0.004)	0.994 (0.006)	0.993 (0.007)
	Sensitivity	0.998 (0.024)	0.972 (0.128)	0.915 (0.263)	0.922 (0.252)
	Specificity	0.999 (0.002)	0.997 (0.004)	0.996 (0.006)	0.994 (0.007)

Table S 10. Average estimated metabolite concentrations for 200 iterations in the sensitivity analysis of different values of σ_ϵ in conjunction with various d values, using shifting variation of ±0.01 ppm. Shaded row shows the results reported in the main text. Corresponding standard deviations are recorded in parentheses

	$\tilde{\boldsymbol{\beta}}$	±0.01 ppm	±0.02 ppm	±0.03 ppm	±0.04 ppm
$1.0\sigma_\epsilon$	$\tilde{\beta}_1$	0.970 (0.106)	0.984 (0.067)	0.968 (0.152)	0.939 (0.228)
	$\tilde{\beta}_2$	0.924 (0.218)	0.953 (0.167)	0.948 (0.184)	0.940 (0.213)
	$\tilde{\beta}_3$	0.903 (0.274)	0.946 (0.201)	0.946 (0.199)	0.926 (0.251)
$1.2\sigma_\epsilon$	$\tilde{\beta}_1$	0.992 (0.028)	0.970 (0.157)	0.962 (0.175)	0.974 (0.141)
	$\tilde{\beta}_2$	0.960 (0.143)	0.973 (0.112)	0.957 (0.169)	0.952 (0.183)
	$\tilde{\beta}_3$	0.953 (0.188)	0.976 (0.128)	0.957 (0.184)	0.949 (0.204)
$1.5\sigma_\epsilon$	$\tilde{\beta}_1$	0.995 (0.015)	0.992 (0.071)	0.947 (0.218)	0.922 (0.263)
	$\tilde{\beta}_2$	0.982 (0.080)	0.990 (0.013)	0.956 (0.179)	0.934 (0.234)
	$\tilde{\beta}_3$	0.984 (0.098)	0.996 (0.009)	0.966 (0.160)	0.936 (0.230)
$1.7\sigma_\epsilon$	$\tilde{\beta}_1$	0.991 (0.071)	0.962 (0.184)	0.894 (0.301)	0.912 (0.279)
	$\tilde{\beta}_2$	0.978 (0.109)	0.988 (0.072)	0.921 (0.250)	0.925 (0.248)
	$\tilde{\beta}_3$	0.984 (0.093)	0.990 (0.071)	0.932 (0.233)	0.925 (0.254)
$2.0\sigma_\epsilon$	$\tilde{\beta}_1$	0.997 (0.009)	0.932 (0.246)	0.848 (0.356)	0.867 (0.336)
	$\tilde{\beta}_2$	0.987 (0.073)	0.972 (0.135)	0.916 (0.266)	0.911 (0.264)
	$\tilde{\beta}_3$	0.989 (0.058)	0.977 (0.128)	0.917 (0.264)	0.922 (0.251)

S4. Additional results and discussion on experimental data analysis

Sample preparation and data acquisition: Three experimental mixtures were prepared with different compositions of 20 amino acids as outlined in Table S11, using 3-(trimethylsilyl) propionic acid (TMSP) as a chemical shift reference at a concentration of 5 mM and 50 mM phosphate buffer in deuterium oxide (D_2O) at pH 7.0 (uncorrected). All solutions were made with D_2O. Each NMR spectrum was collected at 298K using a Bruker AVANCE III-HD 700 MHz spectrometer with 64 scans and 4 dummy scans and a 2 s relaxation delay. The spectra were collected with a spectral width of 11 160 Hz, 32 K data points, and excitation sculpting to suppress the solvent resonance and maintain a flat baseline. NMR spectra were processed using MVAPACK software package. The water signal between 4.65 and 4.9 ppm were removed before outputting data matrices.

Empirically, we collected NMR spectra of 20 individual metabolites (Table S11) at different pH (i.e., 6, 6.5, 7, 7.5, and 8) and temperatures (i.e., 290 K, 294 K, 304 K, and 308 K to obtain the distribution of peak shifting errors for each peak. The process of obtaining the individual NMR spectra was similar to the mixtures' as described above. We observed the absolute maximum amount of shift across all peaks of the 20 metabolites to be 0.01 ppm. Similar to the sensitivity analyses in Section S3.1, we evaluated the performance of the proposed method on the experimental mixtures (as in Section 5.1 of the manuscript) using five different values of d (i.e., 5, 10, 15, 20, and 25). Consistent with the results in Tables S4 and S6, the proposed method performed well at $d = 10$, $d = 15$, and $d = 20$ in terms of accuracy, sensitivity, and specificity, as shown in Table S 12. Therefore, we set $d = 10$ and continued fixing $c_0 = 7\%\text{AUC}$ and $\sigma_0 = \max(y_i^2)/3$ for the real data analysis reported in Section 5 of the manuscript.

To ease the discussion of the performance across various methods regarding accuracy, sensitivity, and specificity using the experimental mixtures in Section 5.1, Table S13 reported the actual number of true positives (TP), false positives (FP), and false negatives (FN). Additionally, multi-panel figures (Fig. S3 – S4) mirrored the observed (black) and reconstructed (red) spectra corresponding to each method using experimental mixture 1 (Table S11). More precisely, Fig. S3 graphically demonstrated our findings discussed in Section 5.1 about the false positive effect caused by Lasso, elastic net, and adaptive Lasso as compared to the proposed method. Similarly, Fig. S4 illustrated both false positive and false negative effects caused by Bayesil, Chenomx, and ASICS in comparison with the proposed method.

Table S 11. List of 20 amino acids used to construct synthetic mixtures

Mixture 1	Mixture 2	Mixture 3	
L-methionine	L-alanine	Glycine	L-leucine
L-serine	L-valine	L-alanine	L-lysine
L-cysteine	L-threonine	L-arginine	L-methionine
L-threonine	L-isoleucine	L-asparagine	L-phenylalanine
L-asparagine	L-leucine	L-asparatic acid	L-proline
L-glutamine	Glycine	L-glutamic acid	L-threonine
	L-proline	L-glutamine	L-tryptophan
		L-histidine	L-tyrosine
		L-isoleucine	L-valine
		L-cysteine	L-serine

Table S 12. Average accuracy, sensitivity, and specificity of the proposed method using 3 experimental mixtures containing 6, 7, and 20 metabolites respectively; and a library size of 61 metabolites at different d values. Shaded column shows the results reported in the main text.

# Met.	Metrics	$d = 5$	$d = 10$	$d = 15$	$d = 20$	$d = 25$
6	Accuracy	0.98	1.00	1.00	0.98	0.98
	Sensitivity	0.86	1.00	1.00	1.00	1.00
	Specificity	1.00	1.00	1.00	0.98	0.98
7	Accuracy	1.00	1.00	1.00	1.00	1.00
	Sensitivity	1.00	1.00	1.00	1.00	1.00
	Specificity	1.00	1.00	1.00	1.00	1.00
20	Accuracy	0.93	0.93	0.92	0.92	0.89
	Sensitivity	1.00	1.00	1.00	1.00	1.00
	Specificity	0.90	0.90	0.88	0.88	0.83

Table S 13. Comparison of proposed method with Lasso, elastic net, adaptive Lasso, Chenomx, Bayesil, and ASICS using 3 experimental mixtures containing 6, 7, and 20 metabolites, respectively; and a library size of 61 metabolites. Total number of true positives (TP), false positives (FP), and false negatives (FN) are recorded

# Met.	Metrics	Proposed method	Lasso	Elastic Net	Adaptive Lasso	Chenomx	Bayesil	ASICS
6	TP	6	6	6	6	4	6	2
	FP	0	7	15	15	10	22	15
	FN	0	0	0	0	2	0	4
7	TP	7	7	7	7	5	7	4
	FP	0	14	22	13	5	22	20
	FN	0	0	0	0	2	0	3
20	TP	20	20	20	20	16	19	9
	FP	4	20	18	19	12	14	39
	FN	0	0	0	0	4	1	11

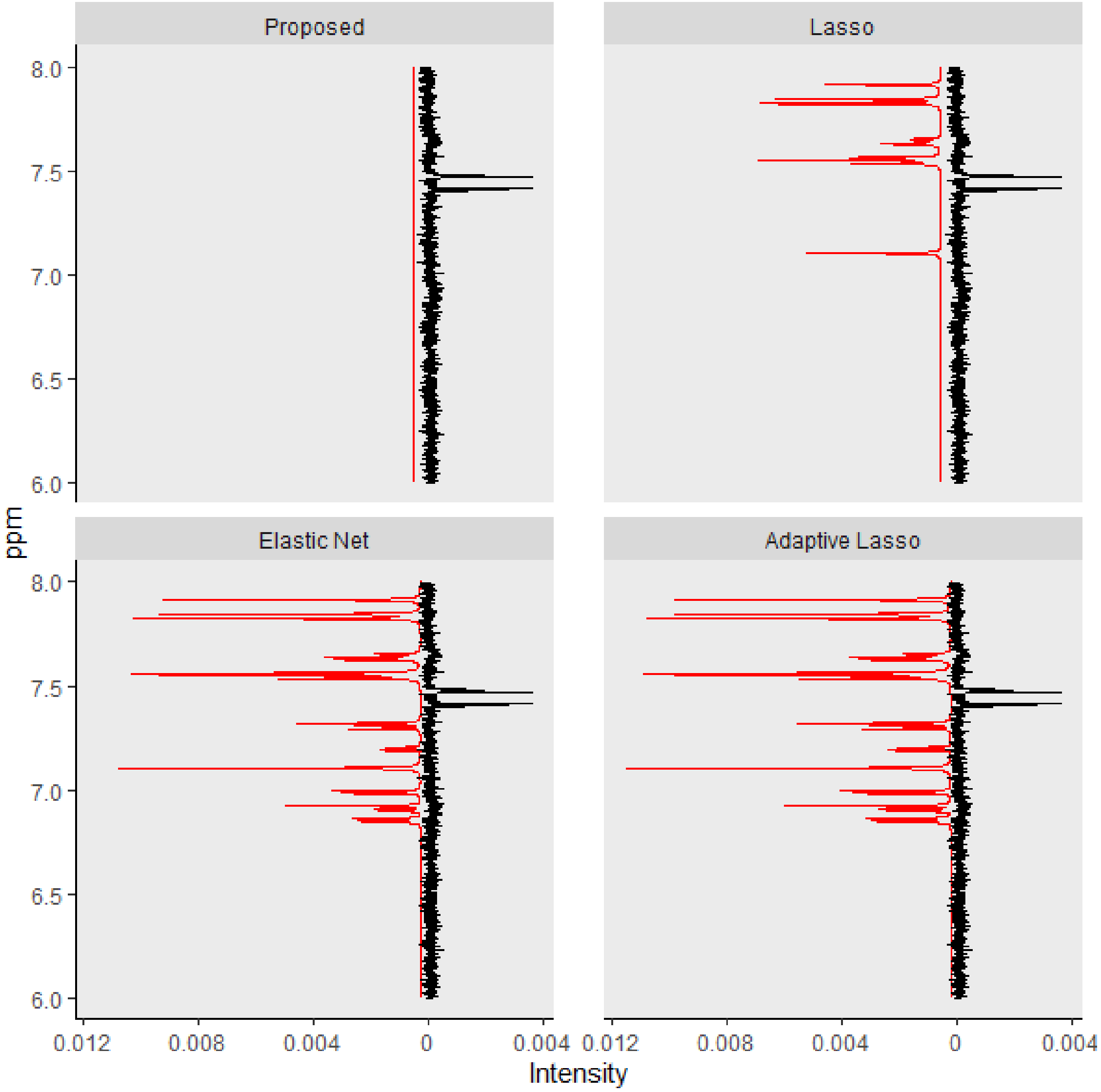

Fig. S3. Zoomed in fitted curve (red) generated from the proposed method, Lasso, elastic net, and adaptive Lasso is mirrored with the observed mixture spectrum (black)

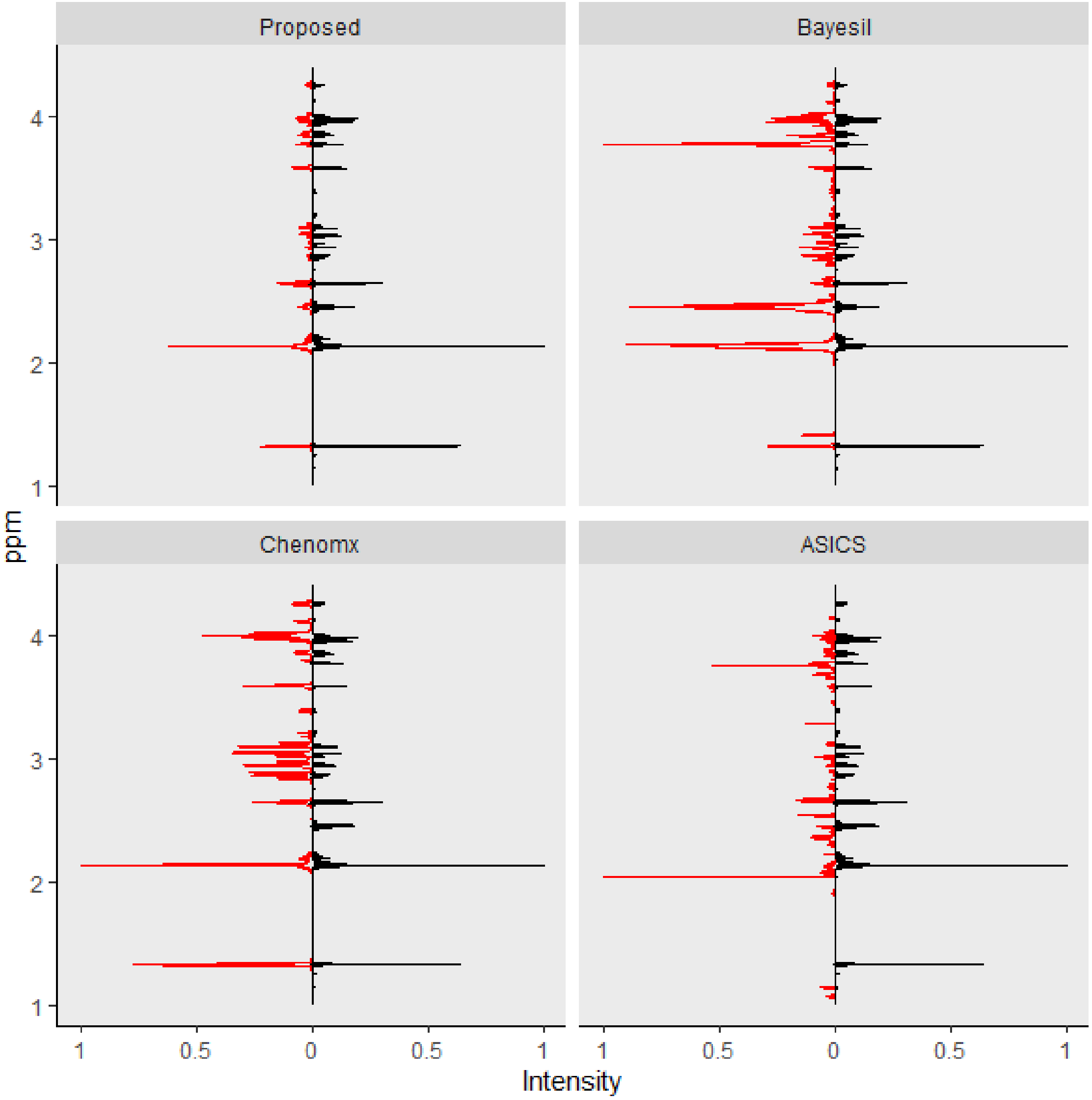

Fig. S4. Zoomed in fitted curve (red) generated from the proposed method, Bayesil, Chenomx, and ASICS is mirrored with the observed mixture spectrum (black)

References

Czanner, Gabriela, Sarma, Sridevi V, Eden, Uri T and Brown, Emery N. (2008). A signal-to-noise ratio estimator for generalized linear model systems. In: *Proceedings of the world congress on engineering*, Volume 2.

Meinshausen, Nicolai and Bühlmann, Peter. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(4), 417–473.