HHS Public Access

Author manuscript

IEEE/ACM Trans Comput Biol Bioinform. Author manuscript; available in PMC 2016 July 01.

Published in final edited form as:

IEEE/ACM Trans Comput Biol Bioinform. 2015 ; 12(4): 914–927. doi:10.1109/TCBB.2014.2377723.

Bayesian Normalization Model for Label-Free Quantitative Analysis by LC-MS

Mohammad R. Nezami Ranjbar,

Department of Electrical and Computer Engineering, Virginia Tech, 900 N. Glebe Road, Arlington, VA 22203, and the Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University, 173 Building D, 4000 Reservoir Road NW, Washington, DC 20057

Mahlet G. Tadesse,

Department of Mathematics and Statistics, Georgetown University, 308 St. Marys Hall, Washington, DC 20057

Yue Wang, and

Department of Electrical and Computer Engineering, Virginia Tech, 900 N. Glebe Road, Arlington, VA 22203

Habtom W. Ressom

Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University, 173 Building D, 4000 Reservoir Road NW, Washington, DC 20057

Mohammad R. Nezami Ranjbar: nranjbar@vt.edu; Mahlet G. Tadesse: mgt26@georgetown.edu; Yue Wang: yuewang@vt.edu; Habtom W. Ressom: hwr@georgetown.edu

Abstract

We introduce a new method for normalization of data acquired by liquid chromatography coupled with mass spectrometry (LC-MS) in label-free differential expression analysis. Normalization of LC-MS data is desired prior to subsequent statistical analysis to adjust variabilities in ion intensities that are not caused by biological differences but experimental bias. There are different sources of bias including variabilities during sample collection and sample storage, poor experimental design, noise, etc. In addition, instrument variability in experiments involving a large number of LC-MS runs leads to a significant drift in intensity measurements. Although various methods have been proposed for normalization of LC-MS data, there is no universally applicable approach. In this paper, we propose a Bayesian normalization model (BNM) that utilizes scan-level information from LC-MS data. Specifically, the proposed method uses peak shapes to model the scan-level data acquired from extracted ion chromatograms (EIC) with parameters considered as a linear mixed effects model. We extended the model into BNM with drift (BNMD) to compensate for the variability in intensity measurements due to long LC-MS runs. We evaluated the performance of our method using synthetic and experimental data. In comparison with several existing methods, the proposed BNM and BNMD yielded significant improvement.

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.

Digital Object Identifier no. 10.1109/TCBB.2014.2377723

Index Terms

Liquid chromatography; mass spectrometry; normalization; bayesian hierarchical model

1 Introduction

Liquid chromatography-mass spectrometry is a promising technology that allows us to measure the abundance of thousands of biomolecules in a sample. Mainly, it is being used to detect differences in the level of abundances of biomolecule in samples from different phenotypes. However there are some computational challenges such as peak detection, alignment, and normalization that continue to be investigated. Specifically, for normalization, the challenge is that either the exact sources of bias are not known or are difficult to be modeled reliably.

Normalization is needed to compensate for differences in sample collection, biomolecule extraction, and instrument variability such as column separation nonlinearity, ionization variability, etc [1] that introduce undesired bias in differential expression analysis.

Analysis of a sample by LC-MS typically generates three pieces of information: a pair of mass-to-charge ratio (m/z) and retention time (RT) along with a related ion intensity. Following processing of data from a set of LC-MS runs, a data matrix is created with each row and column representing a feature (RT, m/z) and a sample, respectively.

LC-MS data processing involves multiple steps including peak detection, deisotoping, peak matching, peak alignment, and intensity normalization. Usually, normalization of the LC-MS data is considered before statistical analysis to remove or decrease the undesired bias [2]. The importance of the sample preparation step to achieve consistent results in different runs of the same experiment was emphasized in recent studies [3] and [4].

Most of the existing normalization methods are applied to the processed data, i.e. the ion intensities obtained by integrating the extracted ion chromotograms. Therefore, the scan-level information is not used for normalization. However, in this study we show that this information is useful for the purpose of modeling bias and performing normalization. On the other hand, one of the most significant sources of bias is analysis order of the runs due to the time it takes for sample preparation, the time samples wait to be analyzed (sample degradation), and the variability of the instrument along the time span of the experiment specifically for experiments involving long queues.

In [5] we proposed a new method for normalization where a stochastic regression approach was utilized to model the variation of intensities across the runs. We also used scan-level information to estimate the variation more accurately. However, by performing normalization for each peak separately, the method is not taking advantage of information from other peaks to adjust the estimates. Moreover, that method heavily relies on the accurate alignment of the scans which is not the case in real situations. In addition, the continuous chromatographic peak is scanned and sampled at discrete times, so that the shift in retention time leads to different points of sampling for peaks from different runs.

Here, we expand the aforementioned work by taking information from other ions into consideration. Specifically, we propose a new Bayesian hierarchical model to quantify LC-MS peak intensities. As our approach does not depend on the perfect alignment of the peaks at scan-level, it can handle peaks with different width in retention time. It can also handle missing peaks or scans by including a an indicator in the model. Finally, we include the variation of the instrument across the runs in the model to address the drift across the queue.

The model includes two main layers. The first layer models the observed intensities based on the original unknown ion abundance, a missing rate for scans, and a peak shape function with several parameters. It also includes the error terms to represent noise in instrument measurements. In the second layer, the parameters of the peak shape function are modeled as a linear combination of several fixed and random effects related to each ion. In addition, a noise term is considered to model the variation of the peak shape across different runs. To learn the parameters of the Bayesian model, Markov chain Monte Carlo (MCMC) is used.

Several studies used Bayesian methods for processing of LC-MS data. Per example, a Bayesian hierarchical method was used for peptide detection in LC-MS proteomics [6]. Other studies such as [7] proposed Bayesian peak detection methods for Matrix-Assisted Laser Desorption/Ionization-Time of Flight (MALDI-TOF) data using a wavelet-based mixed effects model and [8] for LC-MS data using a Gaussian function for peak detection. In a recent work, a Bayesian hierarchical model was used for alignment of LC-MS data [9]. However, to the best of our knowledge, none of these studies used such a model for quantification or modeling the instrument variation using scan-level information as suggested in this work.

2 Methods

We introduce a new hierarchical Bayesian model for normalization that utilizes the scan-level LC-MS data. It can also handle missing scan or noisy intensities as well as misalignment at scan-level. Before introducing the model, we briefly mention several existing methods for LC-MS data normalization. Next, we explain current methods for calculation of peak intensities from LC-MS reads. The following section summarizes several popular peak shape functions used to model chromatographic peaks.

2.1 Existing Normalization Methods

Several methods have been proposed for normalization of LC-MS data. As normalization is a well-known concept in the area of genomics, most of the methods have been adapted from the techniques developed for gene expression microarray data [10], [11], [12], [13], [14]. Usually the underlying assumption of these approaches is that the average biomolecule concentrations should be equal for all samples in the same experiment.

In [15], using the same metabolomic data set employed in this study (Section 3), we reviewed and compared the performance of the following normalization methods: (i) normalization based on total ion count (TIC), (ii) median scale normalization [11], (iii) pretreatment methods [16] such as scaling, centering and transformation, (iv) normalization based on internal standards [17], (v) quantile normalization [14], (vi) MA transform linear

regression normalization [14], (vii) normalization based on quality control (QC) consistency [4], (viii) normalization based on stable features, and (ix) normalization based on analysis order [1]. As a result, we concluded that three methods, normalization based on TIC, median scale normalization, and quantile normalization were consistently outperformed others.

While implementing these methods, we modified or upgraded the algorithms in some cases. Specifically, we introduced a Gaussian process regression model for normalization based on analysis order called 2D-GPRM-EIC [5]. Here, we are expanding the aforementioned work by introducing a more complete model. Therefore, in this paper, we evaluated the performance of our approach with TIC, MedScale, quantile, and 2D-GPRM-EIC normalization methods.

2.2 Chromatographic Peaks

In LC-MS data, each ion or compound is presented by a chromatographic peak. The chromatographic peak is obtained from EIC which is defined by certain range of m/z and retention time specific for each ion. The properties of the chromatographic peak are used for different purposes. For example, the reliability of the measurement can be assessed based on the quality of the chromatographic peak shape. Also the abundance of the related ion is estimated based on the apex or the area under the chromatographic peak.

2.2.1 Intensity Calculation—Most tools such as XCMS [18] and OpenMS [19], calculate the peak intensity based on the peak shape estimate. After peak detection step, either the intensity is calculated as a sum of all scans constructing the peak, or a bell-shaped curve is fitted to the peak and the area under the curve (AUC) or its height is used to represent the peak intensity. However, these two approaches have several shortcomings including the following: (1) in many cases the boundary of the peak is not clearly or precisely defined, (2) ion counts for some scans are missing, and (3) all individual scan reads involve substantial amount of noise. Even for the same peak from the same compound across different replicates, the peak width and the noise level may change leading to a significant change in peak intensity.

2.2.2 Peak Shapes—We used three functions to model chromatographic peak shapes, i.e. Guassian, Gamma, and exponentially modified Gaussian (EMG) from Table 1. Because of the convenience and many interesting properties, Gaussian is a very common choice for many peak detection algorithms, while Gamma has been proposed by few studies [20]. The possible advantage of Gamma over Gaussian is the ability to model asymmetric peaks. Finally, EMG (Fig. 1) is the shape which is suggested by several studies on mass spectrometry because of its goodness of fit [21], [22], [23], [24].

2.3 Bayesian Normalizaiton Model

Here, we explain the data structure followed by a detailed description of the Bayesian hierarchical model, the parameter space, hyperparameters, and the inference procedure.

2.3.1 Data—We considered the model below for scan intensities from detected peaks in the data set:

- $i = 1, \ldots, m$ peaks (ions)
- $j = 1, \ldots, n$ runs (samples)
- $t = 1, \ldots, T_{i,j}$ scans for peak i in run j
- observed intensities $\mathbf{Y} = \{\mathbf{y}_{i,j}\}$

where the number of scans, $T_{i,j}$ for peak i from run j, is not necessarily the same for all peaks. Also as scan time is discrete, this model allows different sampling points along each peak, even for the same ion from different runs. The vector of scan-level intensities from the EIC of peak i from run j is

$$\mathbf{y}_{i,j} = \begin{bmatrix} y_{i,j}(1) & \ldots & y_{i,j}(T_{i,j}) \end{bmatrix}^T \quad (1)$$

2.3.2 Bayesian Hierarchical Model—We modeled the EIC of each peak as

$$y_{i,j}(t) = \eta_{i,j}\gamma_{i,j}(t) f\left(t;\boldsymbol{\phi}_{i,j}\right) + e_{i,j}(t), \quad (2)$$

where η is the ion abundance, $\gamma(t)$ is an indicator random variable to model missing scans, $e_{i,j}(t)$ is random noise, and $f(t)$ is the peak shape function with r parameters summarized in vector $\boldsymbol{\varphi}_{r\times 1}$ and modeled as a combination of some fixed and random effects for each peak across different runs

$$\boldsymbol{\phi}_{i,j} = \boldsymbol{\mu}_i + \mathbf{B}\mathbf{x}_i + \mathbf{A}_i\mathbf{z}_i + \boldsymbol{\varepsilon}_{i,j}. \quad (3)$$

In (3), it is assumed that the peak shape function parameters are linearly dependent on p fixed effects (such as m/z and RT) and q random effects (such as different clusters based on ion annotation). For nonnegative parameters such as variance or scale, we used a log-transformed version. Also we consider an error term for each parameter to include individual variations for every peak in each run. In summary

- $\boldsymbol{\mu}_i$: vector of r mean values of the peak shape parameters
- $\mathbf{x}_i$: vector of p fixed effects
- $\mathbf{z}_i$: binary vector of q random effects
- **B**: Matrix of r fixed effects coefficients

$$\mathbf{B} = \begin{bmatrix} \boldsymbol{\beta}_1^T \\ \vdots \\ \boldsymbol{\beta}_r^T \end{bmatrix} = \begin{bmatrix} \beta_{1,1} & \ldots & \beta_{1,p} \\ \vdots & \ddots & \vdots \\ \beta_{r,1} & \ldots & \beta_{r,p} \end{bmatrix}. \quad (4)$$

- $\mathbf{A}_i$: Matrix of r random effects coefficients

$$\mathbf{A}_i = \begin{bmatrix} \boldsymbol{\alpha}_{i,1}^T \\ \vdots \\ \boldsymbol{\alpha}_{i,r}^T \end{bmatrix} = \begin{bmatrix} \alpha_{i,1,1} & \dots & \alpha_{i,1,q} \\ \vdots & \ddots & \vdots \\ \alpha_{i,r,1} & \dots & \alpha_{i,r,q} \end{bmatrix} \quad (5)$$

which means

$$\phi_{i,j,k} = \mu_{i,k} + \boldsymbol{\beta}_k^T \mathbf{x}_i + \boldsymbol{\alpha}_{i,k}^T \mathbf{z}_i + \varepsilon_{i,j,k} \quad (6)$$

for $i = 1, \dots, m$, $j = 1, \dots, n$, and $k = 1, \dots, r$.

2.3.3 Priors—In this section, we provide the priors for model parameters introduced in the previous section. To make statistical inference easier through MCMC, whenever possible, we selected conjugate priors. Also to take advantage of the Gaussian priors, we sample some of the nonnegative parameters (such as covariance kernel parameters) in the log space [25].

We assume error terms in intensity measurements in (2) are independent random variables generated by a normal distribution

$$e_{i,j}(t) | \sigma_{e_i}^2 \sim \mathcal{N}\left(0, \sigma_{e_i}^2\right), \quad (7)$$

where the noise level for each ion is independent from other ions and its variance follows an inverse-Gamma distribution:

$$\sigma_{e_i}^2 \sim \mathcal{IG}(a_e, b_e). \quad (8)$$

Also for ion abundances, $\eta_{i,j}$:

$$\eta_{i,j} \sim \mathcal{N}\left(\tilde{\eta}_i, \sigma_{\eta_i}^2\right), \quad (9)$$

where:

$$\begin{aligned} \tilde{\eta}_i &\sim \mathcal{N}\left(\eta_0, \sigma_{\eta_0}^2\right) \\ \sigma_{\eta_i}^2 &\sim \mathcal{IG}(a_\eta, b_\eta) \end{aligned} \quad (10)$$

for known hyperparameters a_η and b_η. Moreover, $\eta_0 \approx \hat{\eta_0} = \bar{\eta}$ is estimated from observed data.

For the missing scan indicator variable, a Bernoulli distribution is considered

$$\gamma_{i,j}(t) \sim \mathcal{B}(\lambda_i), \quad (11)$$

where the prior for the missing scan rate, $1 - \lambda_i$, is a Beta distribution:

$$\lambda_i \sim \beta(a_\lambda, b_\lambda), \quad (12)$$

therefore:

$$y_{i,j}(t)|\eta_{i,j}, \lambda_i, \boldsymbol{\phi}_{i,j}, \sigma^2_{e_i} \sim \lambda_i \mathcal{N}\left(\eta_{i,j} f(t;\boldsymbol{\phi}_{i,j}), \sigma^2_{e_i}\right) + (1-\lambda_i)\mathcal{N}\left(0, \sigma^2_{e_i}\right) \quad (13)$$

as $\forall$ *i, j, t* : $y_{i,j}(t) \geq 0$.

Similarly, for error terms in (3)

$$\boldsymbol{\varepsilon}_{ij}|\boldsymbol{\sigma}^2_\varepsilon \sim \mathcal{N}(\mathbf{0}, \sum\nolimits_\varepsilon), \quad (14)$$

where $\sum_\varepsilon = \mathrm{diag}(\boldsymbol{\sigma}^2_\varepsilon)$ and $\boldsymbol{\sigma}^2_\varepsilon = [\ \sigma^2_{\varepsilon_1} \quad \ldots \quad \sigma^2_{\varepsilon_r}\ ]^T$ and for $k = 1, \ldots, r$:

$$\sigma^2_{\varepsilon_k}|a_\varepsilon, b_\varepsilon \sim \mathcal{IG}(a_\varepsilon, b_\varepsilon). \quad (15)$$

Here, we considered the noise terms to be independent. In Section 2.4 we explain how to include the model for variation based on analysis order into the error terms similar to the approach used in [5].

The peak shape parameters are considered as normally distributed

$$\boldsymbol{\phi}_{i,j}|\boldsymbol{\mu}_i, \mathbf{B}, \mathbf{A}_i, \boldsymbol{\sigma}^2_\varepsilon \sim \mathcal{N}(\boldsymbol{\mu}_i + \mathbf{B}\mathbf{x}_i + \mathbf{A}_i\mathbf{z}_i, \sum\nolimits_\varepsilon) \quad (16)$$

where for the mean values

$$\boldsymbol{\mu}_i \sim \mathcal{N}(\boldsymbol{\phi}_0, \sum\nolimits_\mu), \quad (17)$$

where $\boldsymbol{\varphi}_0$ and $\sum_\mu = \mathrm{diag}(\sigma^2_{\mu_1}, \ldots, \sigma^2_{\mu_r})$ are known hyperparameters.

For the fixed effects coefficients ($k = 1, \ldots, r$):

$$\boldsymbol{\beta}_k|\sum\nolimits_\beta \sim \mathcal{N}(\mathbf{0}, \sum\nolimits_\beta), \quad (18)$$

where based on the independence assumption of the coefficients $\boldsymbol{\beta}_k$:

$$\sum\nolimits_\beta = \mathrm{diag}(\sigma^2_\beta) = \begin{bmatrix} \sigma^2_{\beta_1} & 0 & \ldots & 0 \\ 0 & \sigma^2_{\beta_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \ldots & \sigma^2_{\beta_p} \end{bmatrix}_{p \times p}. \quad (19)$$

For scale parameters of the fixed effects coefficient ($\ell = 1, \ldots, p$):

$$\sigma^2_{\beta_\ell}|a_\beta, b_\beta \sim \mathscr{IG}\left(a_\beta, b_\beta\right). \quad (20)$$

Defining $\tau_{\beta_\ell}=\sigma^2_{\beta_\ell}$:

$$\tau_{\beta_\ell} \sim \mathscr{G}\left(a_\beta, b_\beta^{-1}\right). \quad (21)$$

Likewise, for random effects coefficients ($k = 1, \ldots, r$, $\ell = 1, \ldots, q$)

$$\boldsymbol{\alpha}_{i,k}|\sum\nolimits_{\alpha_k} \sim \mathscr{N}\left(\mathbf{0}, \sum\nolimits_{\alpha_k}\right), \quad (22)$$

where dependencies were assumed between random effects:

$$\sum\nolimits_{\alpha_k} = \begin{bmatrix} \sigma^2_{\alpha_k,1} & \ldots & \mathrm{Cov}(\alpha_{k,1}, \alpha_{k,q}) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(\alpha_{k,q}, \alpha_{k,1}) & \ldots & \sigma^2_{\alpha_k,q} \end{bmatrix}$$

and $\boldsymbol{\Sigma}_{a_k,\ell_1,\ell_2} = \rho_{k,\ell_1,\ell_2}\sigma_{a_k,\ell_1}\sigma_{a_k,\ell_2}$ in which ρ_{k,ℓ_1,ℓ_2} is the correlation coefficient between effects ℓ_1 and ℓ_2. For the corresponding covariance matrices ($k = 1, \ldots, r$)

$$\sum\nolimits^{-1}_{\alpha_k}|\boldsymbol{\Psi}_{\alpha_k}, \nu_{\alpha_k} \sim \mathscr{W}_q(\boldsymbol{\Psi}_{\alpha_k}, \nu_{\alpha_k}), \quad (23)$$

where $\nu_{a_k} \in [q, m]$ and:

$$\boldsymbol{\Psi}^{-1}_{\alpha_k} = \nu_{\alpha_k}((1-\omega_k)\hat{\sum}\nolimits_{\alpha_k} + \omega_k(\mathrm{diag}(\hat{\sum}\nolimits_{\alpha_k}))) \quad (24)$$

in which $\hat{\boldsymbol{\Sigma}_{a_k}}$ is a point estimate of $\boldsymbol{\Sigma}_{a_k}$ and $\omega_k \sim \mathscr{U}\left(0, \frac{1}{4}\right)$ is a scalar weight to avoid overestimation of random effects correlations [26]. Here $\mathscr{W}$ and $\mathscr{U}$ are Wishart and uniform probability distributions respectively. Also ν_a is the degrees of freedom for the Wishart distribution which lies uniformly in $[q, m]$. If we select $\nu_a = q$, the prior borrows less information from the data, while choosing $\nu_a = m$ puts the highest weight on the prior from observations. Here, we select $\omega = \frac{1}{4}$ and $\nu_a = q$.

2.3.4 Parameter Space Summary—Summarizing all space parameters in $\boldsymbol{\Theta}$ leads to

$$\boldsymbol{\Theta} = \{\boldsymbol{\eta}, \tilde{\boldsymbol{\eta}}, \sum\nolimits_\eta, \boldsymbol{\Gamma}, \boldsymbol{\lambda}, \sum\nolimits_e, \boldsymbol{\Phi}, \boldsymbol{\mu}, \mathbf{B}, \mathscr{A}, \sum\nolimits_\beta, \sum\nolimits_\alpha, \sum\nolimits_\varepsilon, \boldsymbol{\Psi}_\alpha\},$$

where $\mathbf{B}$, $\boldsymbol{\Sigma}_\beta$, and $\boldsymbol{\Sigma}_\varepsilon$ are defined in (4), (19), and (15) respectively, and

$$\begin{aligned}\boldsymbol{\eta}=\{\eta_{i,j}\},\tilde{\boldsymbol{\eta}} &=\{\tilde{\eta}_i\},\sum_{\eta}=\{\sigma^2_{\eta_i}\},\boldsymbol{\Gamma}=\{\boldsymbol{\gamma}_{i,j}\},\boldsymbol{\lambda}=[\lambda_i]^T,\\ \sum_e &=\{\sigma^2_{e_i}\},\boldsymbol{\Phi}=\{\boldsymbol{\phi}_{i,j}\},\boldsymbol{\mu}=\{\boldsymbol{\mu}_i\},\mathscr{A}=\{\mathbf{A}_i\},\\ \sum_{\alpha} &=\{\sum_{\alpha_k}\},\boldsymbol{\Psi}_{\alpha}=\{\boldsymbol{\Psi}_{\alpha_k}\}.\end{aligned}$$

The posterior probability is

$$P(\boldsymbol{\Theta}|\boldsymbol{Y}) \propto P(\boldsymbol{Y}|\boldsymbol{\Theta})P(\boldsymbol{\Theta}) \quad (25)$$

with the joint distribution of

$$\begin{aligned}P(\boldsymbol{\Theta}) \propto\ & P(\boldsymbol{\eta}|\tilde{\boldsymbol{\eta}},\textstyle\sum_{\eta})P(\tilde{\boldsymbol{\eta}}|\eta_0,\sigma^2_{\eta_0})P(\textstyle\sum_{\eta}|a_{\eta},b_{\eta})\\ \times & P(\boldsymbol{\Gamma}|\lambda)P(\boldsymbol{\lambda}|a_{\lambda},b_{\lambda})P(\textstyle\sum_e|a_e,b_e)P(\boldsymbol{\Phi}|\boldsymbol{\mu},\mathbf{B},\mathscr{A},\textstyle\sum_{\varepsilon})\\ \times & P(\textstyle\sum_{\varepsilon}|a_{\varepsilon},b_{\varepsilon})P(\boldsymbol{\mu}|\boldsymbol{\phi}_0,\textstyle\sum_{\mu})P(\mathbf{B}|\textstyle\sum_{\beta})P(\textstyle\sum_{\beta}|a_{\beta},b_{\beta})\\ \times & P(\mathscr{A}|\textstyle\sum_{\alpha})P(\textstyle\sum_{\alpha}|\boldsymbol{\Psi},\boldsymbol{\nu})P(\boldsymbol{\Psi}|\boldsymbol{\omega}).\end{aligned} \quad (26)$$

2.3.5 Likelihoods—The conditional probabilities can be calculated as $P(\boldsymbol{\Theta}_l|\boldsymbol{\Theta}_{\setminus l})$ given all the other space parameters and hyperparameters. These conditionals will be used to derive the full conditionals.

For the observed data, **Y**, given **Θ**

$$\begin{aligned}P(\boldsymbol{Y}|\boldsymbol{\Theta}) &\propto \prod_{i=1}^{m}\prod_{j=1}^{n}\prod_{t=1}^{T_{i,j}}P(y_{i,j}(t)|\boldsymbol{\Theta})\\ &\propto \prod_{i=1}^{m}\prod_{j=1}^{n}\prod_{t=1}^{T_{i,j}}P(y_{i,j}(t)|\boldsymbol{\phi}_{i,j},\eta_{i,j},\gamma_{i,j}(t),\sigma^2_{e_i}).\end{aligned} \quad (27)$$

The conditional for ion abundances has the following form:

$$P(\boldsymbol{\eta}|\tilde{\boldsymbol{\eta}},\textstyle\sum_{\eta}) \propto \prod_{i=1}^{m}\prod_{j=1}^{n}P(\eta_{i,j}|\tilde{\eta}_i,\sigma^2_{\eta_i}) \quad (28)$$

also for the means, $\tilde{\boldsymbol{\eta}}=[\tilde{\eta_i}]^T$

$$P(\tilde{\boldsymbol{\eta}}|\eta_0,\sigma^2_{\eta_0}) \propto \prod_{i=1}^{m}P(\tilde{\eta}_i|\eta_0,\sigma^2_{\eta_0}) \quad (29)$$

and

$$P(\textstyle\sum_{\eta}|a_{\eta},b_{\eta}) \propto \prod_{i=1}^{m}P(\sigma^2_{\eta_i}|a_{\eta},b_{\eta}). \quad (30)$$

Given missing scan rate and based on the missing at random assumption, we have

$$P(\mathbf{\Gamma}|\boldsymbol{\lambda}) \propto \prod_{i=1}^{m}\prod_{j=1}^{n}\prod_{t=1}^{T_{i,j}} P(\gamma_{i,j}(t)|\lambda_i) \quad (31)$$

for the scan indicator variables, while by knowing hyperparameters

$$P(\boldsymbol{\lambda}|\alpha_\lambda, \beta_\lambda) \propto \prod_{i=1}^{m} P(\lambda_i|\alpha_\lambda, \beta_\lambda). \quad (32)$$

Similarly, for the second layer parameters

$$\begin{aligned} P(\mathbf{\Phi}|\mathbf{\Theta}_{\setminus\Phi}) &\propto P(\mathbf{\Phi}|\boldsymbol{\mu}, \mathbf{B}, \mathscr{A}, \textstyle\sum_\varepsilon) \\ &\propto \prod_{i=1}^{m}\prod_{j=1}^{n} P(\boldsymbol{\phi}_{i,j}|\boldsymbol{\mu}_i, \mathbf{B}, \mathbf{A}_i, \textstyle\sum_\varepsilon) \\ \propto \prod_{i=1}^{m}\prod_{j=1}^{n}\prod_{k=1}^{r} P(\phi_{i,j,k}|\mu_{i,k}, \boldsymbol{\beta}_k, \boldsymbol{\alpha}_{i,k}, \sigma_\varepsilon^2) \end{aligned} \quad (33)$$

and for the related error terms:

$$P(\textstyle\sum_\varepsilon|a_\varepsilon, b_\varepsilon) \propto \prod_{k=1}^{r} P(\sigma_{\varepsilon_k}^2|a_\varepsilon, b_\varepsilon). \quad (34)$$

Finally, for the fixed and random effects coefficients:

$$P(\mathbf{B}|\textstyle\sum_\beta) \propto \prod_{k=1}^{r} P(\boldsymbol{\beta}_k|\textstyle\sum_{\beta_k}) \quad (35)$$

and

$$P(\mathscr{A}|\textstyle\sum_\alpha) \propto \prod_{i=1}^{m} P(\mathbf{A}_i|\textstyle\sum_\alpha) \propto \prod_{i=1}^{m}\prod_{k=1}^{r} P(\boldsymbol{\alpha}_{i,k}|\textstyle\sum_{\alpha_k}). \quad (36)$$

2.3.6 Full Conditionals—Bayes rule was used to find the posterior of each variable $\mathbf{\Theta}_\iota$ given all other variables $\mathbf{\Theta}_{\setminus\iota}$ and observed data, $\mathbf{Y}$

$$\begin{aligned} P(\mathbf{\Theta}_\iota|\boldsymbol{Y}, \mathbf{\Theta}_{\setminus\iota}) &\propto P(\boldsymbol{Y}|\mathbf{\Theta}) P(\mathbf{\Theta}_\iota|\mathbf{\Theta}_{\setminus\iota}) \\ &\propto P(\boldsymbol{Y}|\mathbf{\Theta}) P(\mathbf{\Theta}_{\setminus\iota}|\mathbf{\Theta}_\iota) P(\mathbf{\Theta}_\iota). \end{aligned} \quad (37)$$

In Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2014.2377723, using (37), the final form for all full conditionals are provided in details. Here, we only include the general closed form of the full conditionals based on the hierarchical model.

We begin with the full conditionals for the first layer, starting with ion abundances

$$P(\boldsymbol{\eta}|\boldsymbol{Y},\boldsymbol{\Theta}_{\setminus\eta}) \propto P(\boldsymbol{Y}|\boldsymbol{\Phi},\boldsymbol{\eta},\boldsymbol{\Gamma},\textstyle\sum_e)P(\boldsymbol{\eta}|\tilde{\boldsymbol{\eta}},\textstyle\sum_\eta)$$
$$\propto \prod_{i=1}^{m}\prod_{j=1}^{n}\left(P(\eta_{i,j}|\tilde{\eta}_i,\sigma^2_{\eta_i}) \times \prod_{t=1}^{T_{ij}} P\left(\mathbf{y}_{i,j}(t)|\boldsymbol{\phi}_{i,j},\eta_{i,j},\gamma_{i,j}(t),\sigma^2_{e_i}\right)\right) \quad (38)$$

also for mean abundances, $\tilde{\boldsymbol{\eta}}$:

$$P(\tilde{\boldsymbol{\eta}}|\boldsymbol{Y},\boldsymbol{\Theta}_{\setminus\tilde{\eta}}) \propto P(\boldsymbol{Y}|\boldsymbol{\Phi},\boldsymbol{\eta},\boldsymbol{\Gamma},\textstyle\sum_e)P(\boldsymbol{\eta}|\tilde{\boldsymbol{\eta}},\textstyle\sum_\eta)P(\tilde{\boldsymbol{\eta}}|\eta_0,\sigma^2_{\eta_0})$$
$$\propto \prod_{i=1}^{m}\left(P(\tilde{\eta}_i|\eta_0,\sigma^2_{\eta_0})\prod_{j=1}^{n}P(\eta_{i,j}|\tilde{\eta}_i,\sigma^2_{\eta_i})\right) \quad (39)$$

and the variance of the abundances, $\boldsymbol{\Sigma}_{\eta}$:

$$P(\textstyle\sum_\eta|\boldsymbol{Y},\boldsymbol{\Theta}_{\setminus\sigma^2_\eta}) \propto P(\boldsymbol{Y}|\boldsymbol{\Phi},\boldsymbol{\eta},\boldsymbol{\Gamma},\textstyle\sum_e)P(\boldsymbol{\eta}|\tilde{\boldsymbol{\eta}},\textstyle\sum_\eta)P(\textstyle\sum_\eta|a_\eta,b_\eta)$$
$$\propto \prod_{i=1}^{m}\left(P(\sigma^2_{\eta_i}|a_\eta,b_\eta)\prod_{j=1}^{n}P(\eta_{i,j}|\tilde{\eta}_i,\sigma^2_{\eta_i})\right). \quad (40)$$

For indicating variables modeling the missing scans

$$P(\boldsymbol{\Gamma}|\boldsymbol{Y},\boldsymbol{\Theta}_{\setminus\Gamma}) \propto P(\boldsymbol{Y}|\boldsymbol{\Phi},\boldsymbol{\eta},\boldsymbol{\Gamma},\textstyle\sum_e)P(\boldsymbol{\Gamma}|\boldsymbol{\lambda})$$
$$\propto \prod_{i=1}^{m}\prod_{j=1}^{n}\prod_{t=1}^{T_{i,j}}P(y_{i,j}(t)|\boldsymbol{\phi}_{i,j},\eta_{i,j},\gamma_{i,j}(t),\sigma^2_{e_i})P(\gamma_{i,j}(t)|\lambda_i) \quad (41)$$

and similarly, for the missing rate of the scans:

$$P(\boldsymbol{\lambda}|\boldsymbol{Y},\boldsymbol{\Theta}_{\setminus\lambda}) \propto P(\boldsymbol{Y}|\boldsymbol{\Phi},\boldsymbol{\eta},\boldsymbol{\Gamma},\textstyle\sum_e)P(\boldsymbol{\lambda}|\boldsymbol{\Gamma})$$
$$\propto P(\boldsymbol{\Gamma}|\boldsymbol{\lambda})P(\boldsymbol{\lambda}|a_\lambda,b_\lambda)$$
$$\propto \prod_{i=1}^{m}\left(P(\lambda_i|a_\lambda,b_\lambda)\prod_{j=1}^{n}\prod_{t=1}^{T_{i,j}}P(\gamma_{i,j}(t)|\lambda_i)\right). \quad (42)$$

The intensity error terms in layer one from (2) have a full conditional in the form of

$$P(\textstyle\sum_e|\boldsymbol{Y},\boldsymbol{\Theta}_{\setminus\sigma^2_e}) \propto P(\boldsymbol{Y}|\boldsymbol{\Phi},\boldsymbol{\eta},\boldsymbol{\Gamma},\textstyle\sum_e)P(\textstyle\sum_e|a_e,b_e)\prod_{i=1}^{m}\left(P(\sigma^2_{e_i}|a_e,b_e)\prod_{j=1}^{n}\prod_{t=1}^{T_{ij}}P(y_{ij}(t)|\boldsymbol{\phi}_{ij},\eta_{ij},\gamma_{ij}(t),\sigma^2_{e_i})\right). \quad (43)$$

Similarly, for the peak shape function parameters, i.e. the second layer

$$P(\boldsymbol{\Phi}|\boldsymbol{Y},\boldsymbol{\Theta}_{\setminus\Phi}) \propto P(\boldsymbol{Y}|\boldsymbol{\Theta})P(\boldsymbol{\Phi}|\boldsymbol{\Theta}_{\setminus\Phi})$$
$$\propto \prod_{i=1}^{m}\prod_{j=1}^{n}\left(P(\boldsymbol{\phi}_{i,j}|\boldsymbol{\mu}_i,\mathbf{B},\mathbf{A}_i,\textstyle\sum_\varepsilon) \times \prod_{t=1}^{T_{i,j}}P(y_{i,j}(t)|\boldsymbol{\phi}_{i,j},\eta_{i,j},\gamma_{i,j}(t),\sigma^2_{e_i})\right) \quad (44)$$

and the error terms:

$$P(\sum_{\varepsilon}|\boldsymbol{Y},\boldsymbol{\Theta}_{\setminus\sum_{\varepsilon}}) \propto P(\boldsymbol{\Phi}|\boldsymbol{\mu},\mathbf{B},\mathscr{A},\sum_{\varepsilon})P(\sum_{\varepsilon}|a_{\varepsilon},b_{\varepsilon})$$
$$\propto \prod_{k=1}^{r}\left(P(\sigma^2_{\varepsilon_k}|a_{\varepsilon},b_{\varepsilon})\prod_{i=1}^{m}\prod_{j=1}^{n}P(\phi_{i,j,k}|\mu_{i,k},\boldsymbol{\beta}_k,\boldsymbol{\alpha}_{i,k},\sigma^2_{\varepsilon_k})\right). \quad (45)$$

For the mean parameters the assumption is that they are independent for each ion and each peak shape parameter

$$P(\boldsymbol{\mu}|\boldsymbol{Y},\boldsymbol{\Theta}_{\setminus\mu}) \propto P(\boldsymbol{\Phi}|\boldsymbol{\mu},\mathbf{B},\mathscr{A},\sum_{\varepsilon})P(\boldsymbol{\mu}|\boldsymbol{\phi}_0,\sum_{\mu})$$
$$\propto \prod_{i=1}^{m}\left(P(\boldsymbol{\mu}_i|\boldsymbol{\phi}_0,\sum_{\mu})\prod_{j=1}^{n}P(\boldsymbol{\phi}_{i,j}|\boldsymbol{\mu}_i,\mathbf{B},\mathbf{A}_i,\sum_{\varepsilon})\right). \quad (46)$$

Fixed effects coefficients were also assumed to be i.i.d. normally distributed

$$P(\mathbf{B}|\boldsymbol{Y},\boldsymbol{\Theta}_{\setminus B}) \propto P(\boldsymbol{\Phi}|\boldsymbol{\mu},\mathbf{B},\mathscr{A},\sum_{\varepsilon})P(\mathbf{B}|\sum_{\beta})$$
$$\propto \prod_{i=1}^{m}\prod_{j=1}^{n}P(\boldsymbol{\phi}_{i,j}|\boldsymbol{\mu}_i,\mathbf{B},\mathbf{A}_i,\sum_{\varepsilon})\prod_{k=1}^{r}P(\boldsymbol{\beta}_k|\sum_{\beta}) \quad (47)$$

and their covariance matrices

$$P(\sum_{\beta}|\boldsymbol{Y},\boldsymbol{\Theta}_{\setminus\sum_{\beta}}) \propto P(\mathbf{B}|\sum_{\beta})P(\sum_{\beta}|a_{\beta},b_{\beta})$$
$$\propto \prod_{k=1}^{r}P(\boldsymbol{\beta}_k|\sum_{\beta})P(\sum_{\beta}|a_{\beta},b_{\beta}). \quad (48)$$

Likewise, random effects coefficients were considered to be independent for each peak shape parameter, but correlated across different ions

$$P(\mathscr{A}|\boldsymbol{Y},\boldsymbol{\Theta}_{\setminus\mathscr{A}}) \propto P(\boldsymbol{\Phi}|\boldsymbol{\mu},\mathbf{B},\mathscr{A},\sum_{\varepsilon})P(\mathscr{A}|\sum_{\alpha})$$
$$\propto \prod_{i=1}^{m}\left(\prod_{j=1}^{n}P(\boldsymbol{\phi}_{i,j}|\boldsymbol{\mu}_i,\mathbf{B},\mathbf{A}_i,\sum_{\varepsilon})\prod_{k=1}^{r}P(\boldsymbol{\alpha}_{i,k}|\sum_{\alpha_k})\right) \quad (49)$$

so for related covariance matrices we have

$$P(\sum_{\alpha}|\boldsymbol{Y},\boldsymbol{\Theta}_{\setminus\sum_{\alpha}}) \propto P(\mathscr{A}|\sum_{\alpha})P(\sum_{\alpha}|\boldsymbol{\Psi}_{\alpha},\boldsymbol{\nu})$$
$$\propto \prod_{k=1}^{r}\left(P(\sum_{\alpha_k}|\nu_k,\boldsymbol{\Psi}_{\alpha_k})\prod_{i=1}^{m}P(\boldsymbol{\alpha}_{i,k}|\sum_{\alpha_k})\right). \quad (50)$$

2.4 Bayesian Normalization Model with Drift

So far we showed how to use a Bayesian model to quantify the abundances of the ions in an LC-MS experiment by borrowing information from other ions across all runs. Previously, we used Gaussian process regression to model the variation of the instrument based on analysis order which is reflected in the intensity of ions across runs [27]. Following that, we extended the idea to utilize scan-level information [5].

Here, we expand the idea to include the instrument variability in the Bayesian hierarchical model. To do so, we assume the error terms in peak shape function parameters are generated by a Gaussian process. In other words, we assume for peak i across runs $j = 1, \ldots, n$ and for peak shape function parameter φ_k as in (3):

$$\begin{bmatrix} \varepsilon_{i,1,k} & \ldots & \varepsilon_{i,n,k} \end{bmatrix}^T | \sum_{\varepsilon_k} \sim \mathcal{N}(0, \sum_{\varepsilon_k}), \quad (51)$$

where $\boldsymbol{\Sigma}_{\varepsilon_k}$ is an $n \times n$ matrix and for runs j_1 and j_2 we have

$$\sum_{\varepsilon_k, j_1, j_2} = \sigma^2_{\varepsilon_k} h(|j_1 - j_2|; \mathbf{c}). \quad (52)$$

Here $h(j_1, j_2, \mathbf{c})$ is a valid covariance kernel function with parameters $\mathbf{c}$, for example, a Matern kernel function [28] as used in [27]. Therefore:

$$\sum_{\varepsilon_k} = \sigma^2_{\varepsilon_k} \mathbf{H}, \quad (53)$$

where:

$$\mathbf{H} = \mathbf{H}(\mathbf{c}) = \begin{bmatrix} h(1,1;\mathbf{c}) & \ldots & h(1,n;\mathbf{c}) \\ \vdots & \ddots & \vdots \\ h(n,1;\mathbf{c}) & \ldots & h(n,n;\mathbf{c}) \end{bmatrix}_{n \times n}. \quad (54)$$

2.4.1 Full Conditionals with Drift—From (27) and (51), we can derive the full conditional for $\boldsymbol{\Phi}$ as

$$\begin{aligned} P(\boldsymbol{\Phi} | \boldsymbol{Y}, \boldsymbol{\Theta}_{\setminus \Phi}) &\propto P(\boldsymbol{Y} | \boldsymbol{\Theta}) P(\boldsymbol{\Phi} | \boldsymbol{\Theta}_{\setminus \Phi}) \\ \propto \prod_{i=1}^{m} \left(\prod_{k=1}^{r} P(\boldsymbol{\phi}_{i,k} | \mu_{i,k}, \boldsymbol{\beta}_k, \boldsymbol{\alpha}_{i,k}, \sum_{\varepsilon_k}) . \times \prod_{j=1}^{n} \prod_{t=1}^{T_{i,j}} P(y_{i,j}(t) | \boldsymbol{\phi}_{i,j}, \eta_{i,j}, \gamma_{i,j}(t), \sigma^2_{e_i}) \right) \end{aligned} \quad (55)$$

where $\boldsymbol{\phi}_{i,k} = [\varphi_{i,1,k}, \ldots, \varphi_{1,n,k}]^T$ and $\boldsymbol{\varphi}_{i,j} = [\varphi_{i,j,1}, \ldots, \varphi_{i,j,r}]^T$.

Consequently, it is required to update the full conditionals for some the second layer parameters. While Appendix B, available in the online supplemental material, provides the details on how to update (45), (46), (47), and (49), the general form of the full conditionals are included here.

Beginning with peak shape function parameters error covariance, $\boldsymbol{\Sigma}_{\varepsilon} = \{\boldsymbol{\Sigma}_{\varepsilon_k}\}$

$$\begin{aligned} P(\sum_{\varepsilon} | \boldsymbol{Y}, \boldsymbol{\Theta}_{\setminus \sum_{\varepsilon}}) &\propto P(\boldsymbol{Y} | \boldsymbol{\Theta}) P(\boldsymbol{\Theta}_{\setminus \sum_{\varepsilon}} | \sum_{\varepsilon}) P(\sum_{\varepsilon}) \\ \propto \prod_{k=1}^{r} \left(P(\sum_{\varepsilon_k} | a_{\varepsilon}, b_{\varepsilon}, \mathbf{c}) \prod_{i=1}^{m} P(\boldsymbol{\phi}_{i,k} | \mu_{i,k}, \boldsymbol{\beta}_k, \boldsymbol{\alpha}_{i,k}, \sum_{\varepsilon_k}) \right). \end{aligned} \quad (56)$$

For mean parameters, $\boldsymbol{\mu}$, we have

$$P(\mu|\boldsymbol{Y},\boldsymbol{\Theta}_{\setminus\mu}) \propto P(\boldsymbol{\Phi}|\boldsymbol{\mu},\mathbf{B},\mathscr{A},\textstyle\sum_{\varepsilon})P(\boldsymbol{\mu}|\phi_0,\textstyle\sum_{\mu})$$
$$\propto \prod_{i=1}^{m}\left(P(\boldsymbol{\mu}_i|\phi_0,\textstyle\sum_{\mu})\prod_{k=1}^{r}P(\boldsymbol{\phi}_{i,k}|\mu_{i,k},\boldsymbol{\beta}_k,\boldsymbol{\alpha}_{i,k},\textstyle\sum_{\varepsilon_k})\right). \quad (57)$$

For fixed effects coefficients

$$P(\mathbf{B}|\boldsymbol{Y},\boldsymbol{\Theta}_{\setminus B}) \propto P(\boldsymbol{\Phi}|\boldsymbol{\mu},\mathbf{B},\mathscr{A},\textstyle\sum_{\varepsilon})P(\mathbf{B}|\textstyle\sum_{\beta})$$
$$\propto \prod_{k=1}^{r}\left(P(\boldsymbol{\beta}_k|\textstyle\sum_{\beta})\prod_{i=1}^{m}P(\boldsymbol{\phi}_{i,k}|\mu_{i,k},\boldsymbol{\beta}_k,\boldsymbol{\alpha}_{i,k},\textstyle\sum_{\varepsilon_k})\right). \quad (58)$$

Likewise, for random effects coefficients

$$P(\mathscr{A}|\boldsymbol{Y},\boldsymbol{\Theta}_{\setminus\mathscr{A}}) \propto P(\boldsymbol{\Phi}|\boldsymbol{\mu},\mathbf{B},\mathscr{A},\textstyle\sum_{\varepsilon})P(\mathscr{A}|\textstyle\sum_{\alpha})$$
$$\propto \prod_{i=1}^{m}\prod_{k=1}^{r}\left(P(\boldsymbol{\phi}_{i,k}|\mu_{i,k},\boldsymbol{\beta}_k,\boldsymbol{\alpha}_{i,k},\textstyle\sum_{\varepsilon_k})P(\boldsymbol{\alpha}_{i,k}|\textstyle\sum_{\alpha_k})\right). \quad (59)$$

2.5 MCMC Sampling

MCMC was used to infer the model space parameters. As shown in Algorithm 1, for the parameters $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_{N_G}$ with a known posterior density function, we used Gibbs sampling. For the rest of parameters $\boldsymbol{\Theta}_{N_G+1} = \mathrm{Y}_1, \ldots, \boldsymbol{\Theta}_{N_G+N_{\mathrm{MH}}} = \mathrm{Y}_{N_{\mathrm{MH}}}$, we utilized Metropolis-Hastings update by using a proposal distribution $Q_{\mathrm{J}}(\mathrm{Y}_{\mathrm{J}})$.

For the proposal distributions, Q, we used multivariate Gaussian centered at the value of the variable from previous iteration. Also the variance of the proposal distribution was considered to be a diagonal identity matrix.

As mentioned in Section 2.3, Appendices A and B provide all full conditionals. Based on that, we used Gibbs sampling for all space parameters except $\boldsymbol{\varphi}$, $\boldsymbol{\phi}$, and $\mathbf{c}$, where Metropolis-Hastings was used. Thus, each MCMC iteration, includes one update for each space parameter. Although, for several space parameters, we need to include other loops for $i = 1, \ldots, m$ ions, $j = 1, \ldots, n$, $k = 1, \ldots, r$ peak shape function parameters, and $t = 1, \ldots, T_{i,j}$ scans in case required.

Algorithm 1

Inference by MCMC

Require: hyperparameters

 # Initialization #

for $i = 1$ to m **do**

  for $j = 1$ to n **do**

    find $\hat{\boldsymbol{\varphi}}_{i,j}$ by curve fitting

  end for

end for

Estimate $\hat{\boldsymbol{\mu}}$, $\hat{\mathbf{B}}$, $\hat{\mathscr{A}}$, and $\hat{\boldsymbol{\Sigma}}_{\varepsilon}$ by LME

Calculate point estimates of the related statistics

Set $\boldsymbol{\Theta}^{(1)} = \left\{ \boldsymbol{\Theta}_1^{(1)}, \ldots, \boldsymbol{\Theta}_{N_G}^{(1)}, \boldsymbol{\Upsilon}_1^{(1)}, \ldots, \boldsymbol{\Upsilon}_{N_{MH}}^{(1)} \right\}$

 # MCMC sampling #

for $\wp = 2$ to $N_\wp$ **do**

 # Gibbs sampling for $\boldsymbol{\Theta}_1$ to $\boldsymbol{\Theta}_{N_G}$

$$\boldsymbol{\Theta}_1^{(\wp+1)} \sim P\left(\boldsymbol{\Theta}_1 \mid \mathbf{Y}, \boldsymbol{\Theta}_2^{(\wp)}, \boldsymbol{\Theta}_3^{(\wp)}, \ldots, \boldsymbol{\Theta}_{N_G}^{(\wp)}, \boldsymbol{\Upsilon}^{(\wp)}\right)$$
$$\boldsymbol{\Theta}_2^{(\wp+1)} \sim P\left(\boldsymbol{\Theta}_2 \mid \mathbf{Y}, \boldsymbol{\Theta}_1^{(\wp+1)}, \boldsymbol{\Theta}_3^{(\wp)}, \ldots, \boldsymbol{\Theta}_{N_G}^{(\wp)}, \boldsymbol{\Upsilon}^{(\wp)}\right)$$
$$\vdots$$
$$\boldsymbol{\Theta}_{N_G}^{(\wp+1)} \sim P\left(\boldsymbol{\Theta}_{N_G} \mid \mathbf{Y}, \boldsymbol{\Theta}_1^{(\wp+1)}, \ldots, \boldsymbol{\Theta}_{N_G-1}^{(\wp+1)}, \boldsymbol{\Upsilon}^{(\wp)}\right)$$

 # Metropolis–Hastings for $\mathbf{Y}_1$ to $\mathbf{Y}_{N_{MH}}$

for $\jmath = 1$ to N_{MH} **do**

$$\boldsymbol{\Upsilon}_\jmath^{\bigstar} \sim Q_\jmath\left(\boldsymbol{\Upsilon}_\jmath \mid \mathbf{Y}, \boldsymbol{\Theta}_1^{(\wp)}, \ldots, \boldsymbol{\Theta}_{N_G}^{(\wp)}, \boldsymbol{\Upsilon}_{\setminus\jmath}^{(\wp)}\right)$$
$$r_\jmath = \min\left(1, \frac{P(\boldsymbol{\Upsilon}_\jmath^{\bigstar}) Q_\jmath(\boldsymbol{\Upsilon}_\jmath^{(\wp)}; \boldsymbol{\Upsilon}_\jmath^{\bigstar})}{P(\boldsymbol{\Upsilon}_\jmath^{(\wp)}) Q_\jmath(\boldsymbol{\Upsilon}_\jmath^{\bigstar}; \boldsymbol{\Upsilon}_\jmath^{(\wp)})}\right)$$
$$R \sim \mathcal{U}(0,1)$$
$$\boldsymbol{\Upsilon}_\jmath^{(\wp+1)} = \begin{cases} \boldsymbol{\Upsilon}_\jmath^{(\wp+1)} & R \le r_\jmath \\ \boldsymbol{\Upsilon}_\jmath^{(\wp)} & R > r_\jmath \end{cases}$$

end for

end for

2.5.1 Initialization Using Linear Mixed Effects (LME)—Before running MCMC, the initial values for some parameters including peak shape function parameters were found by using a curve fitting estimator. We selected the Levenberg-Marquardt algorithm [29] to find the parameter values for the curve fitting problem as a nonlinear least-squares optimization. Then, by using a LME model, the initial values for the parameters in the second layer of the Bayesian model were estimated.

Therefore, for each peak shape, first, we fit the curve to use the optimized parameters for every ion across the runs. Then, the required statistics such as means and variances are calculated based on these estimated parameters. This includes $\hat{\boldsymbol{\varphi}}_{i,j}$ which can be used to derive $\hat{\boldsymbol{\mu}}$, $\hat{\mathbf{B}}$, $\hat{\mathcal{A}}$, and $\hat{\boldsymbol{\Sigma}}_\varepsilon$ using (3). Also kusing (2), we can estimate other parameters such as $\hat{\eta}_{i,j}$, $\hat{\sigma}_e^2$, and $\hat{\lambda}_i$ and their corresponding statistics required for initialization of the MCMC.

3 Results and Discussion

To evaluate the performance of our approach, we used synthetic and experimental data sets. In the former case, as we have the ground truth, the performance assessment is more accurate. In the second case, we used QC runs and internal standards as a reference by assuming small technical variability in the sample preparation step.

3.1 Data

3.1.1 Synthetic Data Set—We used equations (2) and (3) to generate a simulated data set to test our approach before applying the method to real data sets. To do this, we considered different signal to noise levels for error terms in the first and second layers of our Bayesian model. We also considered different peak shapes. More details are provided in Appendix C, available in the online supplemental material.

To generate the data set, we assumed that peak shape parameters are linearly dependent on m/z and RT with small slopes, i.e. $p = 2$ in (4). Also m/z and RT values are randomly drawn in the mass-time space covering a range of 50–600Da and 60–480sec, respectively. Peak abundances were randomly selected based on a empirical distribution obtained from a typical experimental study.

Also, three random effects, i.e. $q = 3$ in (5), and four groups of ions with equal number of members were considered. The first three groups are affected only by the one random effect respectively, while the fourth group has dependencies on both the first and the second random effects.

Moreover, three different average scan rates per peak, 15, 20, and 30 are considered. The reason behind this consideration is that the effective signal to noise ratio is a descending function of the peak width (for more details, see Appendix C, available in the online supplemental material).

To simulate real conditions, we added random noise to the generated peaks at scan-level. Also we added noise to the peak shape function parameters to account for instrument variability. We used three levels of SNR for the second layer parameters, 20, 25, and 30 dB denoted as SNR_2. For the first layer, 40, 45, and 50 dB denoted as SNR_1 were considered regarding the effect of scan rate mentioned above. Therefore SNR_1 is based on the ratio of η to σ_e in (2) and SNR_2 is based on the ratio of μ to σ_ε in (3).

Fig. 3 shows the effect of noisy parameters on the simulated QC runs. The EICs were generated assuming the EMG peak shape, while using the same abundance for each peak. However, by adding noise to the parameters, we can see a notable change in the observed intensity calculated based the area under the curve as illustrated in Fig. 3 the related coefficient of variation (CV).

3.1.2 Experimental Data Sets—We have two experimental data sets. The first data set is from a metabolomic experiment introduced in [30]. It includes 89 runs of experimental samples and 20 QC runs pooled from the those samples. The samples were run in a randomized order and in two ionization modes (positive and negative). Also five internal

standards have been spiked in all runs which can be utilized to evaluate the effect of normalization methods.

To use this data set, it was needed to retrieve the EICs from the raw data. For this purpose, we used XCMS with the Regions Of Interest (ROI) option [31] extract scan-level peak intensities as well as the mass, retention time, and ion annotation information. In total, using SNR of 30, there are 802 and 537 ions (peaks or features) detected for positive and negative modes respectively, leading to 87,418 and 58,533 EICs with an average of 13 scans per EIC. Following that, we performed alignment at peak level, to make sure that we retrieve the peaks of the same ion across different runs. We used our script in R to export EICs for all peaks using some XCMS functions.

In addition, we used CAMERA [32] to obtain the attributes of the ion groups as monoisotopes, isotopes, adducts, and adducts-isotopes. In addition, we log-transformed the data to make it amenable to equations (9) and (10). Finally, similar to the synthetic data set, we used m/z and RT as fixed effects and ion attributes as random effects for this data set. As mass and retention time ranges may vary in different experiments, here, normalized m/z and RT values were used, by applying a linear function to map all the values in [0, 1] interval.

The second experimental data set is from a proteomic experiment using the same set that includes 14 QC runs injected in between experimental samples that were run in one batch.

MaxQuant [33] was used to process the data providing a list of 2,123 peptides for all runs. We filtered the peptides based on their presence in all samples reducing the number to 1,014. Based on the estimated retention times from Max-Quant, EICs were obtained by our in-house scripts from the raw data following conversion into mzMXL format was used for this purpose. The obtained data set thus includes 14,196 EICs with an average of 29 scans per EIC from the MS1 level. Similar to the metabolomic data set, m/z of and RT values were mapped in [0, 1] interval.

3.2 Evaluation Approach

While ground truth was used for the synthetic data, internal standards and QC runs employed to evaluate the performance of the methods on the experimental data set. Four measures were used for evaluation:

3.2.1—The most popular measure, coefficient of variation of internal standards or ion intensities of QC runs comparing before and after normalization:

$$CV_i = \frac{\overline{\sigma}_i}{\overline{\eta}_i} \quad (60)$$

where:

$$\overline{\eta}_i = \frac{1}{n}\sum_{j=1}^{n}\eta_{i,j},\ \ \overline{\sigma}_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(\eta_{i,j}-\overline{\eta}_i)^2}. \quad (61)$$

3.2.2—Decrease in median standard deviation (MSD) of QC runs obtained from all ions [1].

3.2.3—Number of ions with statistically significant variation (NISV) for QC runs which is calculated by a one-way repeated-measures ANOVA model based on the experimental design (considering batch and group effects together with technical replicates) to estimate the variability of QCs across runs:

$$\eta_{ijk}^{QC}=\mu_i+\alpha_{ij}+\zeta_{ik}+\varepsilon_{ijk}, \quad (62)$$

where η_{ijk}^{QC} is the intensity of kth QC run for ith ion in batch j and ζ is the random effect with $E_k[\zeta_{ik}] = 0$. A normalization method is evaluated on the basis of the number of ions with reduced variance of ζ_{ik}. We evaluate this by using the F test for the ratio of the sum of squares from ζ to the sum of the squares of ε which is the unexplained variation or error.

3.2.4—Relative mean squared error (RMSE) for estimation of the abundances of each ion:

$$RMSE=\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{\hat{\eta}_{i,j}-\eta_{i,j}}{\eta_{i,j}}\right)^2. \quad (63)$$

As mentioned in Section 2.1, we compared our method with existing normalization methods studied in [15] including TIC, MedScale, and quantile normalization, as well as the 2D Gaussian process regression normalization introduced in [5].

3.3 Performance Evaluation

3.3.1 Computational Performance—A workstation with 16 cores and 128 GB of memory was used for the analysis. The run times for simulated, metabolomic, and proteomic data sets were 12, 27, and 15 hrs, respectively. We used diagnostic tests to investigate the convergence of the MCMC procedure. We discarded the first 5,000 MCMC samples as burn-in and estimated the parameters of interest based on the remaining 10,000 samples. We also checked the mixing of the MCMC chains to avoid very high or very low rejection rates.

3.3.2 Evaluation via Synthetic Data Set—As described in previous sections, we used the synthetic data generated with several different sets of parameters including SNR levels for both layers and scan rates. Table 2 shows the results for different levels of SNR_1 and SNR_2 when changing the scan rate. The RMSE is calculated based on (63).

It can be seen that by increasing the SNR in the first layer (measured intensities), the error was reduced for almost all scan rates at each SNR level for the second layer (measurement parameters). Also, in general, EMG peak shape performs better compared to Gaussian and Gamma while Gaussian provides less error compared to the Gamma. The reason is that EMG is more appropriate to model the asymmetric chromatographic peaks while it is also able to model the symmetric peaks with Gaussian peak shapes. On the other hand, although Gamma is able to model the asymmetric peaks better than Gaussian, but in noisy conditions,

it may not model the asymmetry good enough when using fewer parameters compared to EMG.

Another observation is that BNMD is more efficient for SNR_2 of 25 dB compared to higher or lower levels. This is because for lower levels, the parameters are too noisy and the method loses efficiency to model the drift and correct for it. Also for higher levels, the effect of noise on parameters is less significant and as a consequence, there is less bias to be addressed.

In Table 3, using equation (60), the CV of QC samples is shown to compare different normalization methods with the proposed method in this study. As mentioned in Section 2.1, these methods were selected based on [5] and [15]. It can be seen that the CV is decreasing by increasing SNR at the intensity level as expected. Also by taking scan rates into consideration, it is observed that the average error is lower for 20 scans per peak for most of the cases. Finally, if we increase SNR at the parameter level, we see the same trend.

3.3.3 Evaluation via Experimental Data sets—Table 4 shows the RMSE for five internal standards spiked in all runs from the metabolomic data set. The average estimated abundance of the standards were used as the approximate true abundance. Similar to what is suggested by results from synthetic data, it can be seen that EMG provided a better performance compared to Gaussian and Gamma respectively. Fig. 5 depicts two examples of fitting EMG and Gaussian peaks to real EICs from experimental data. As these EICs are asymmetric, EMG provides a better fit. Also compared to raw data, BNM and BNMD were able to reduce the RMSE for all of the standards.

Based on (62), we found the number of ions showing statistically significant variation in their intensities from QC runs. Table 5 shows the percentage of such ions. As illustrated, we compared the proposed method with other existing methods, where BNM and BNMD were able to decrease the number of ions with significant variation in QC runs more than any other method. In the same table, we also included the decrease in MSD of QCs for different normalization methods, where BNM and BNMD were able to outperform others.

Table 6 shows the estimated parameters for the fixed and random effects in our Bayesian hierarchical model. In the table, the values for different parameters are provided, including the average and the standard deviation. As illustrated, there is a noticeable relationship between mass over charge and peak apex location, width, and skewness. However there is no strong evidence for dependency of first parameter on the retention times. Also, it can be seen that the peaks are wider in average for later elution times. Although, peak skewness slightly changes by increasing RT. In addition, correlation coefficients are provided based on covariance matrixes of random effects. Based on that, the random effects of first and the second groups show rather stronger correlation compared to others.

Fig. 6 shows the the intensity of two internal standards before and after normalization. It can be seen that BNM is able to reduce the variation across the runs, however BNMD is successful to capture the overall variation based on analysis order and correct for it to some extent.

We can see a similar trend when evaluating the performance of the methods on the proteomic data. Table 7 shows the average in CV of ions for QC runs after applying different normalization for the proteomic data set. The numbers in parentheses shows the standard deviation. As illustrated, BNM and BNMD were able to reduce the variability in QC runs for this experimental data set more efficiently compared to other methods. Also, evaluation of different methods in terms of the decrease in MSD reveals a noticeable improvement when using the proposed methods.

Finally, Fig. 7 shows the distribution of the intensities before and after normalization comparing different methods for the proteomic data set. As the data set only includes QCs, we expect to see a consistent intensity across runs. As illustrated, BNM and BNMD were more successful to decrease the CV of detected ions in average.

4 CONCLUSION

We proposed a new normalization method for analysis of LC-MS data from label-free experiments. The method utilizes a Bayesian hierarchical model for accurate quantification of peak intensities by using scan-level information and measurements from all measured ions. Also by using Gaussian process regression, we expanded the model to include the drift based on analysis order. In addition, the model is able to handle noisy scan-level data and address missing ion counts when the measurements come from the background noise. Moreover, it does not rely on perfect alignment of the peaks at scan-level.

Using synthetic and experimental LC-MS data, we demonstrated that our model outperforms existing normalization methods. We also used internal standards and QC runs as a reference for comparisons. In addition, we showed that the proposed drift model can improve the estimation of the ion abundances.

We also showed that assuming an exponentially modified Gaussian peak shape function leads to a better performance in terms of RMSE of estimated abundances as well as decrease in intensity variations of QC runs.

Although many of the assumed parameters are learned through MCMC, the proposed model requires specification of several hyperparameters which rely on properties of the raw data. In addition, working with data sets involving either very small sample size or few number of ions can reduce the efficiency of the method.

One potential direction to extend the proposed approach is to integrate normalization with other steps in the data processing pipeline. For example if alignment is merged with normalization, it can improve the peak detection leading to a decrease in the intensity estimation error.

One can also consider different priors or likelihoods for the Bayesian hierarchical model. For example, instead of a normal distribution, a Pareto-log-normal density may be used for the mean values of ion abundances [34]. Also we may include correlations between peak shape function parameters by putting more constraints on their covariance matrix.

Regarding the variables with the Metropolis-Hastings update, a better proposal distribution may be used. Several methods have been introduced to perform adaptive updates. For instance, a local multivariate normal distribution can be utilized for this purpose by using gradient and Hessian of the original posterior distribution function. Also nonrandom Hamiltonian updates can be considered.

Finally, as the algorithm is computationally intensive, it is possible to use parallel processing in the implementation to speed up the execution and reduce the run time. This can be done by updating independent variables in parallel loops.

Acknowledgments

This work was supported by the National Cancer Institute Grant R01CA143420.

References

1. Kultima K, Nilsson A, Scholz B, Rossbach UL, Flth M, Andrn PE. Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides. Mol Cell Proteomics. 2009; 8(10):2285–2295. [PubMed: 19596695]
2. Listgarten J, Emili A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. Mol Cell Proteomics. 2005; 4:419–434. [PubMed: 15741312]
3. Tuli L, Ressom HW. LC-MS based detection of differential protein expression. J Proteomics Bioinformat. 2009; 2(10):416–438.
4. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIn-tyre S, Anderson N, Brown M, Knowles JD, Halsall A, Haselden JN, Nicholls AW, Wilson ID, Kell DB, Goodacre R. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protocols. Jun; 2011 6(7):1060–1083. [PubMed: 21720319]
5. Nezami Ranjbar MR, Zhao Y, Tadesse MG, Wang Y, Ressom HW. Gaussian process regression model for normalization of LC-MS data using scan-level information. Proteome Sci. 2013; 11(Suppl 1):S13. [PubMed: 24564985]
6. Sun Y, Zhang J, Braga-Neto U, Dougherty ER. BPDA2da 2D global optimization-based bayesian peptide detection algorithm for liquid chromatograph–mass spectrometry. Bioinformatics. 2012; 28(4):564–572. [PubMed: 22155863]
7. Morris JS, Brown PJ, Herrick RC, Baggerly KA, Coombes KR. Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. Biometrics. 2008; 64(2):479–489. [PubMed: 17888041]
8. Viv-Truyols G. Bayesian approach for peak detection in two-dimensional chromatography. Anal Chem. 2012; 84(6):2622–2630. [PubMed: 22229801]
9. Tsai TH, Tadesse MG, Wang Y, Ressom HW. Profile-based LC-MS data alignment—A Bayesian approach. IEEE/ACM Trans Comput Biol Bioinformat. Mar-Apr;2013 10(2):494–503.
10. Callister SJ, Barry RC, Adkins JN, Johnson ET, Qian WJ, Webb-Robertson BJM, Smith RD, Lipton MS. Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. J Proteome Res. Feb; 2006 5(2):277–286. [PubMed: 16457593]
11. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. Feb.2002 30(4):e15. [PubMed: 11842121]
12. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics. Jul; 2002 18(suppl 1):S96–S104. [PubMed: 12169536]

13. Anderle M, Roy S, Lin H, Becker C, Joho K. Quantifying reproducibility for differential proteomics: Noise analysis for protein liquid chromatography-mass spectrometry of human serum. Bioinformatics. Dec; 2004 20(18):3575–3582. [PubMed: 15284095]
14. Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. Jan; 2003 19(2): 185–193. [PubMed: 12538238]
15. Nezami Ranjbar, MR.; Zhao, Y.; Tadesse, MG.; Wang, Y.; Ressom, HW. Evaluation of normalization methods for analysis of LC-MS data. Proc. IEEE Int. Conf. Bioinformat. Biomed. Workshops; 2012; p. 610-617.
16. van den Berg R, Hoefsloot H, Westerhuis J, Smilde A, van der Werf M. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. BMC Genomics. Jun.2006 7(1):142. [PubMed: 16762068]
17. Sysi-Aho M, Katajamaa M, Yetukuri L, Oresic M. Normalization method for metabolomics data using optimal selection of multiple internal standards. BMC Bioinformatics. 2007; 8(1):93. [PubMed: 17362505]
18. Smith CA, Want EJ, OMaille G, Abagyan R, Siuzdak G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem. 2006; 78(3):779–787. [PubMed: 16448051]
19. Sturm M, Bertsch A, Gropl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K, Kohlbacher O. OpenMS–An open-source software framework for mass spectrometry. BMC Bioinformat. 2008; 9(1):163.
20. Golubev A. Exponentially modified Gaussian relevance to distributions related to cell proliferation and differentiation. J Theor Biol. 2010; 262(2):257–266. [PubMed: 19825376]
21. Grushka E. Characterization of exponentially modified Gaussian peaks in chromatography. Anal Chem. 1972; 44(11):1733–1738. [PubMed: 22324584]
22. Foley JP, Dorsey JG. A review of the exponentially modified Gaussian function: Evaluation and subsequent calculation of universal data. J Chromatographic Sci. 1984; 22(1):40–46.
23. Naish P, Hartwell S. Exponentially modified Gaussian functionsa good model for chromatographic peaks in isocratic HPLC. Chromatographia. 1988; 26(1):285–296.
24. Kalambet Y, Kozmin Y, Mikhailova K, Nagaev I, Tikhonov P. Reconstruction of chromatographic peaks using the exponentially modified Gaussian function. J Chem. 2011; 25(7):352–356.
25. Barber, D.; Cemgil, AT.; Chiappa, S. Bayesian Time Series Models. Cambridge, U.K: Cambridge Univ. Press; 2011.
26. Brooks, S.; Gelman, A.; Jones, G.; Meng, X-L. Handbook of Markov Chain Monte Carlo. Boca Raton, FL, USA: CRC Press; 2011.
27. Nezami Ranjbar, MR.; Tadesse, MG.; Wang, Y.; Ressom, HW. Normalization of LC-MS data using Gaussian process. Proc. IEEE Int. Workshop Genomic Signal Process. Stat; 2012; p. 187-190.
28. Rasmussen, CE.; Williams, CKI. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning Series). Cambridge, MA, USA: MIT Press; Nov. 2005
29. Moré, JJ. Numerical Analysis. New York, NY, USA: Springer; 1978. The levenberg-marquardt algorithm: Implementation and theory; p. 105-116.
30. Xiao JF, Varghese RS, Zhou B, Nezami Ranjbar MR, Zhao Y, Tsai T-H, Di Poto C, Wang J, Goerlitz D, Luo Y, Cheema AK, Sarhan N, Soliman H, Tadesse MG, Ziada DH, Ressom HW. LCMS based serum metabolomics for identification of hepatocellular carcinoma biomarkers in Egyptian cohort. J Proteome Res. 2012; 11(12):5914–5923. [PubMed: 23078175]
31. Benton HP, Wong DM, Trauger SA, Siuzdak G. XCMS2: Processing tandem mass spectrometry data for metabolite identification and structural characterization. Anal Chem. 2008; 80(16):6382–6389. [PubMed: 18627180]
32. Kuhl C, Tautenhahn R, Boettcher C, Larson TR, Neumann S. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal Chem. 2012; 84:283–289. [PubMed: 22111785]
33. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008; 26(12):1367–1372. [PubMed: 19029910]

34. Lu C, King RD. An investigation into the population abundance distribution of mRNAs, proteins, and metabolites in biological systems. Bioinformatics. 2009; 25(16):2020–2027. [PubMed: 19535531]

Biographies

Mohammad R. Nezami Ranjbar is currently working toward the PhD degree in the Department of Electrical and Computer Engineering at Virginia Tech. He is also a research assistant at the Lombardi Comprehensive Cancer Center, Georgetown University. His research focuses on applications of statistical machine learning to problems in omics data analysis.

Mahlet G. Tadesse is an associate professor in the Department of Mathematics and Statistics at Georgetown University. Her research focuses on the development of statistical and computational tools for the analysis of large-scale genomic data. She is particularly interested in stochastic search methods and Bayesian inferential strategies to identify structures and relationships in high-dimensional data sets.

Yue Wang is the endowed Grant A. Dove professor of electrical and computer engineering at Virginia Tech. His research interests focus on statistical pattern recognition, machine learning, signal and image processing, with applications to computational bioinformatics and biomedical imaging for human disease research.

Habtom W. Ressom is a professor in the Department of Oncology and the director of the Genomics and Epigenomics Shared Resource at the Lombardi Comprehensive Cancer Center, Georgetown University. His research is focused on the application of statistical and machine learning methods for analysis of high-dimensional omics data. He is a senior member of the IEEE.

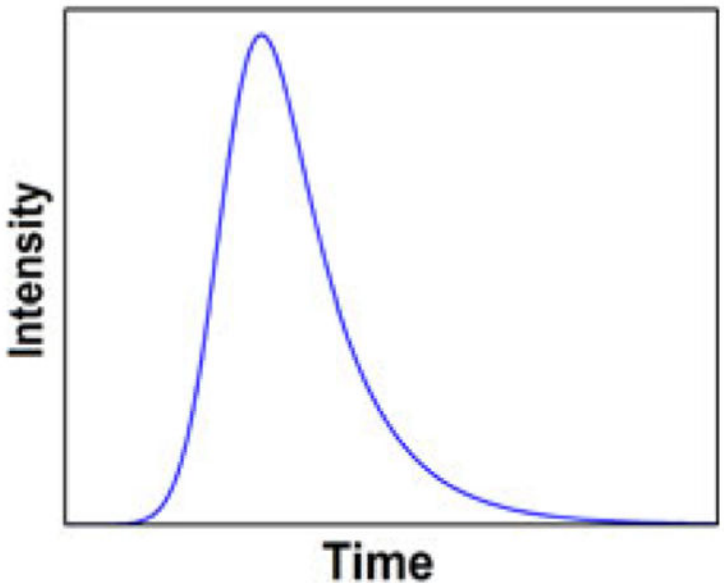

Fig. 1.
Exponentially modified Gaussian.

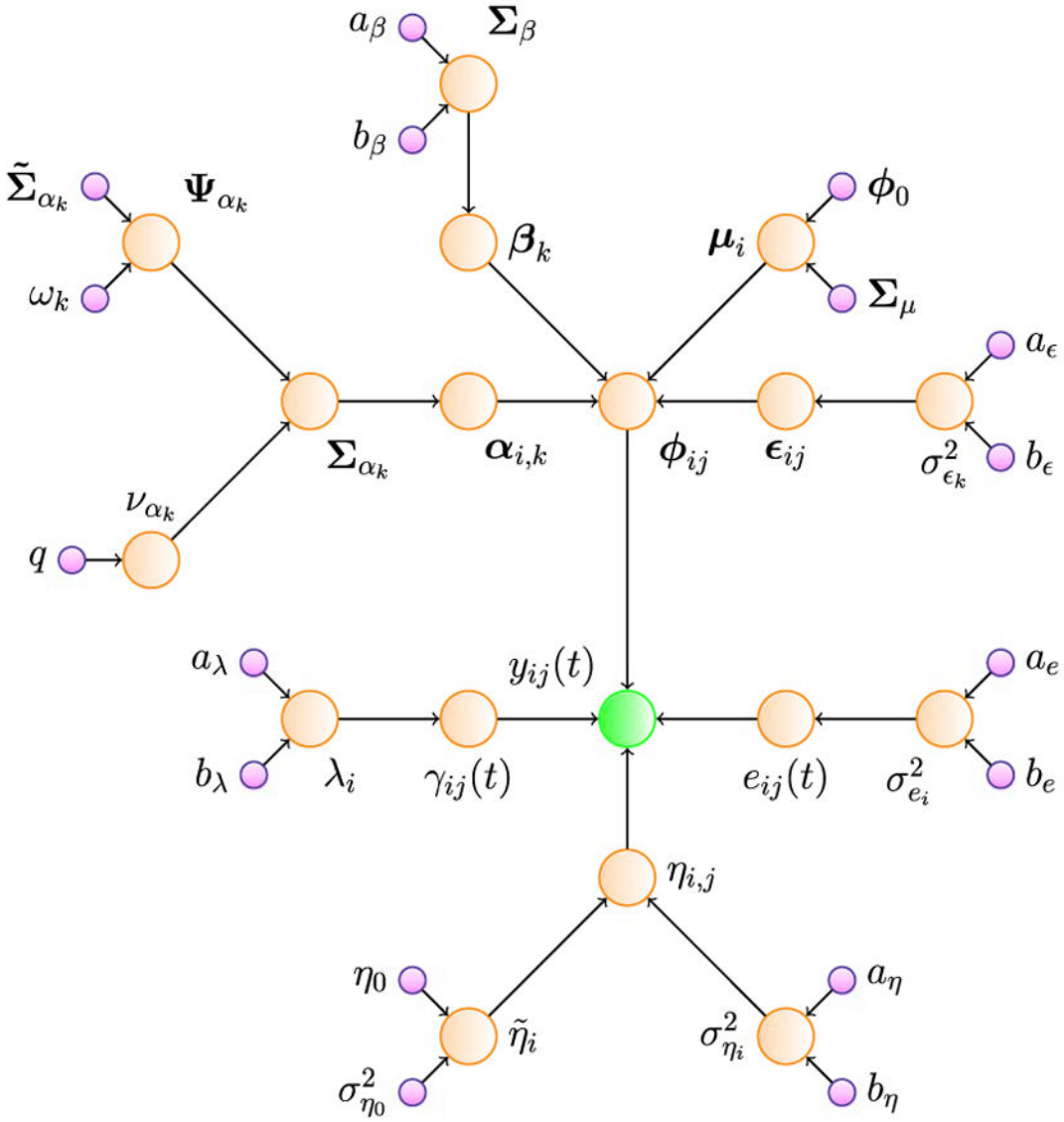

Fig. 2.
The Bayesian hierarchical model network of space parameters and hyperparameters.

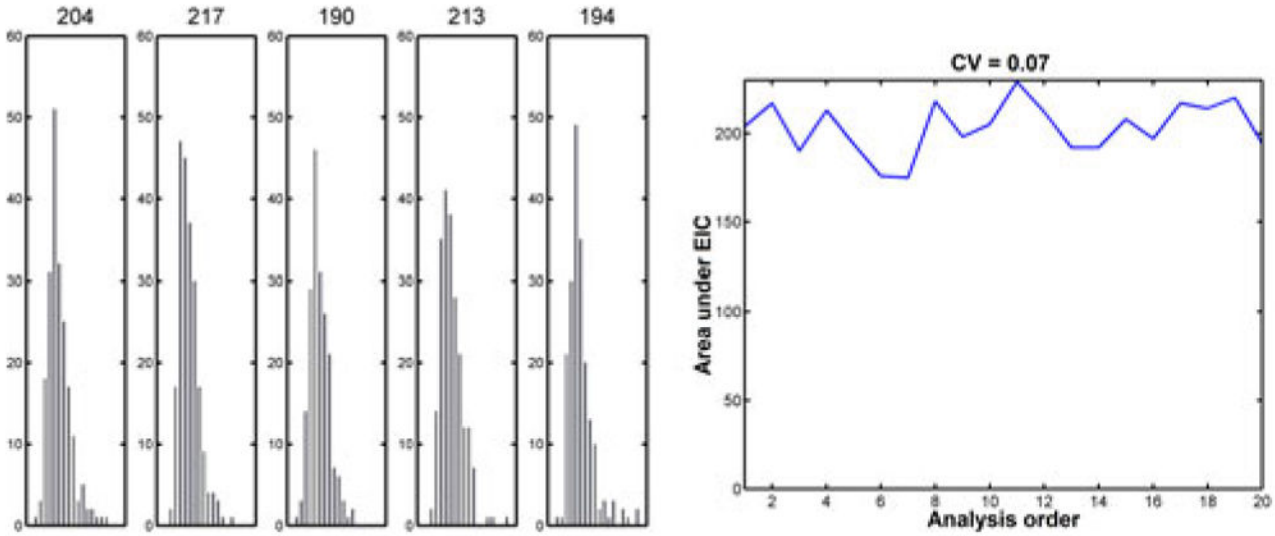

Fig. 3.

(left) Five peaks from the same ion in simulated QC runs assuming EMG peak shape with true abundance of 200. The numbers above each box are the estimated intensity based on sum of the ion counts. (right) The intensity for 20 peaks from the same ion where the CV is 7%.

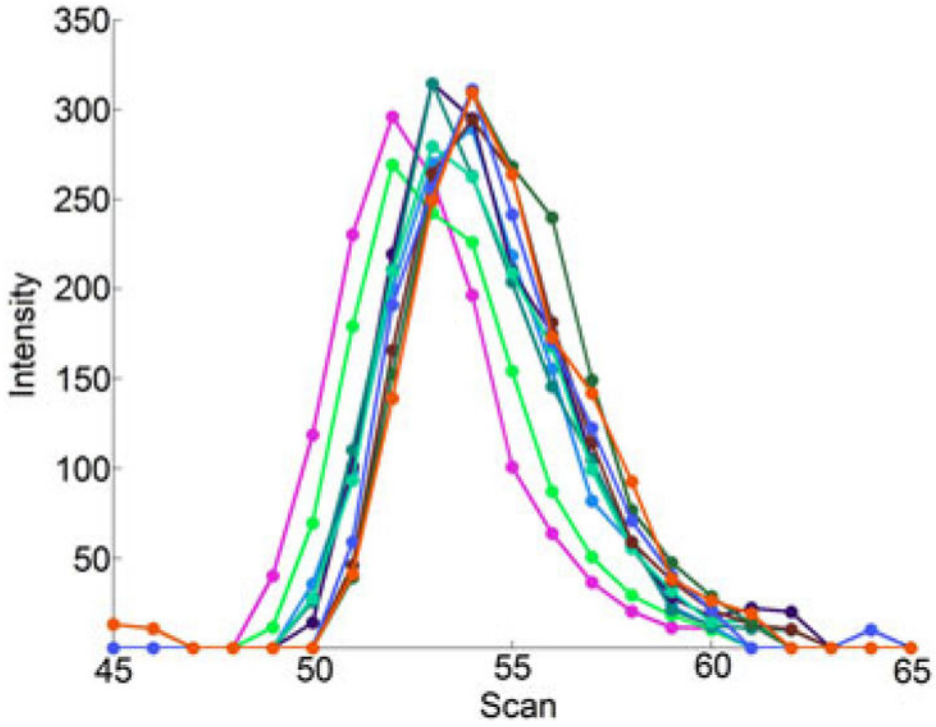

Fig. 4.
EICs of the same ion from the experimental data showing misalignment at scan-level.

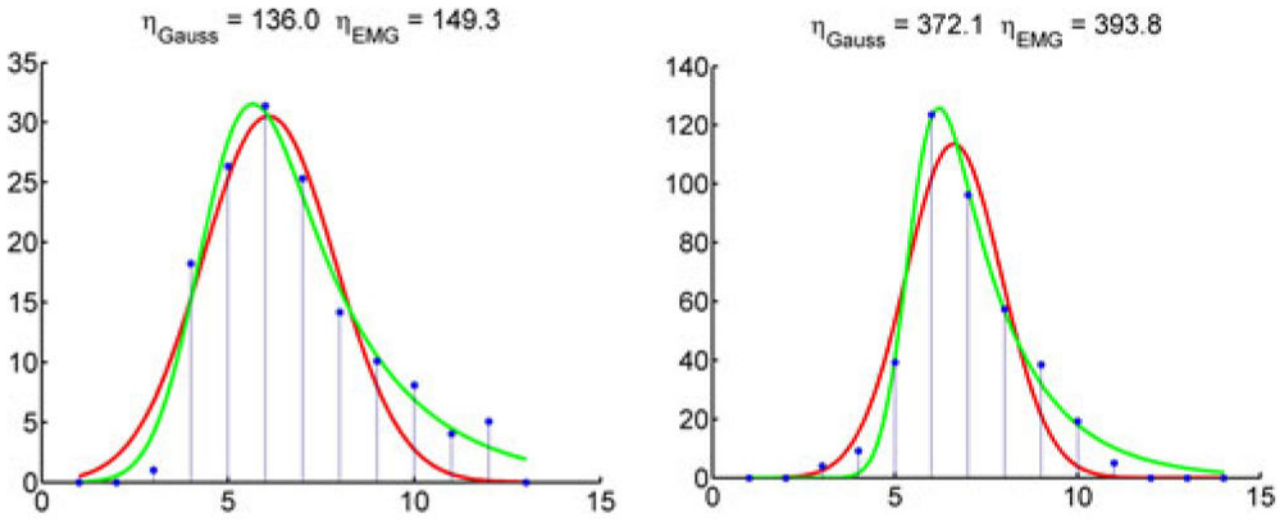

Fig. 5.

Example of Gaussian (dark) and EMG (light) peak shapes fitted to two experimental EICs with the estimated abundance, η, for each case.

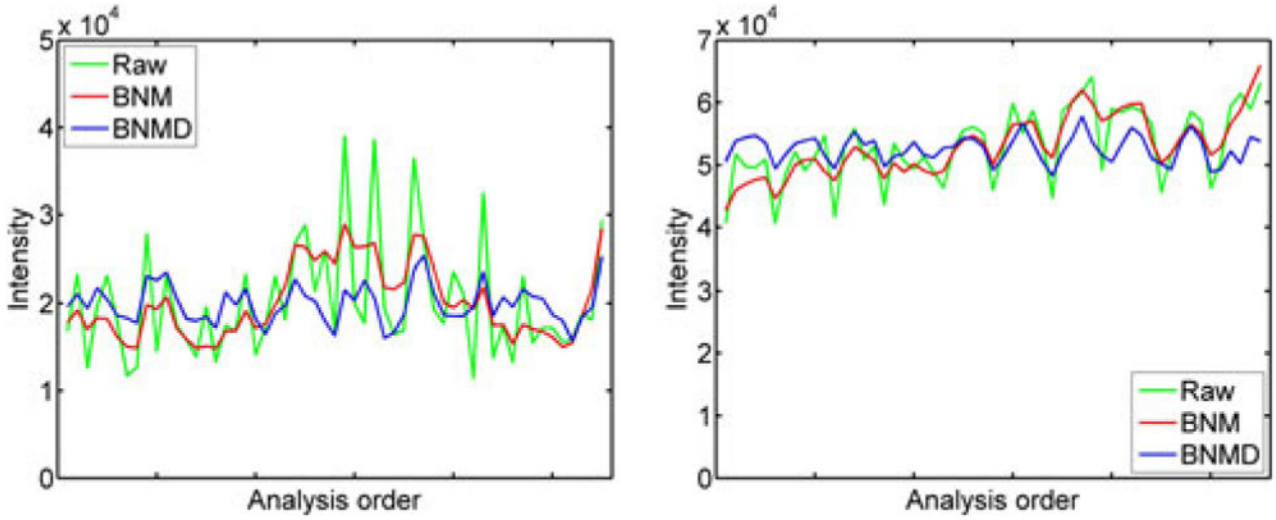

Fig. 6.

Intensities from two internal standards in metabolomic data set from raw data (green), normalized by BNM (red), and normalized by BNMD (blue).

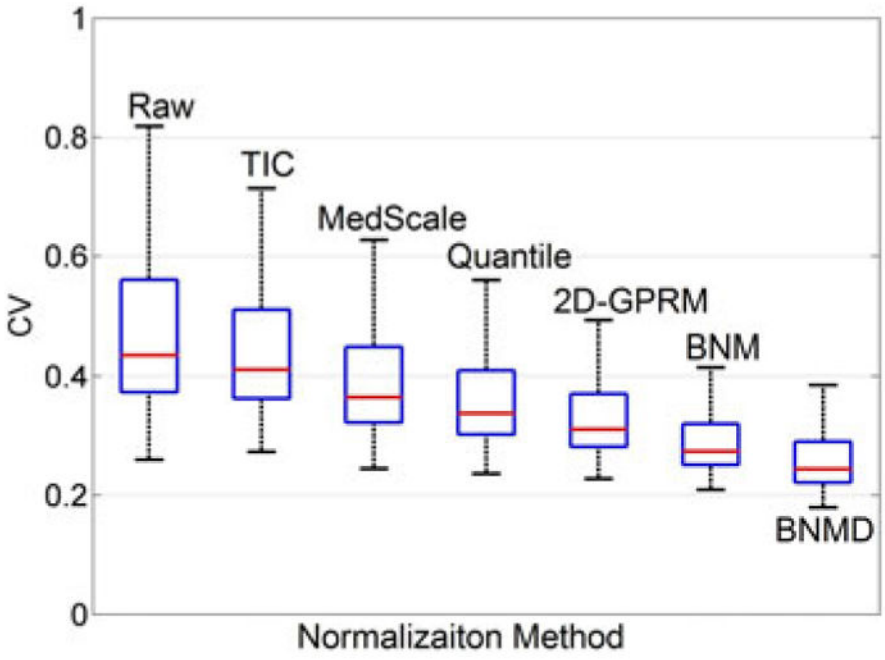

Fig. 7.
CVs of ions comparing different normalization methods for the proteomic data set.

TABLE 1

Chromatographic Peak Shape Functions and Their Parameters

Function	***f*(*t*)**
Gaussian	$(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(t-\mu)^2\right)$
Gamma	$(\Gamma(v))^{-1}\zeta^{-k}t^{v-1}\exp(-\frac{1}{\zeta}t)$
Exponentially	$\frac{1}{2}\zeta\exp\left(\frac{1}{2}\zeta(2\mu+\zeta\sigma^2-2t)\right)$
Modified Gaussian	$\times\left(1-erf(\frac{1}{\sqrt{2}\sigma}(\mu+\zeta\sigma^2-t))\right)$

TABLE 2

RMSE for Synthetic Data Before (Raw) and After Normalization by BNM and BNMD

SNR_1 (dB)	Shape	Rate	RMSE (Original) SNR_2 (dB)			RMSE (BNM) SNR_2 (dB)			RMSE (BNMD) SNR_2 (dB)		
			20	25	30	20	25	30	20	25	30
40	Gaussian	15	0.321	0.237	0.153	0.281	0.183	0.129	0.262	0.147	0.106
		20	0.298	0.224	0.146	0.272	0.175	0.118	0.250	0.139	0.096
		30	0.310	0.235	0.149	0.283	0.182	0.131	0.259	0.144	0.101
	Gamma	15	0.349	0.245	0.168	0.293	0.195	0.138	0.267	0.150	0.110
		20	0.345	0.240	0.152	0.285	0.187	0.131	0.264	0.142	0.103
		30	0.352	0.246	0.170	0.286	0.191	0.142	0.270	0.149	0.111
	EMG	15	0.294	0.223	0.134	0.260	0.176	0.122	0.239	0.131	0.098
		20	0.281	0.216	0.126	0.258	0.153	0.109	0.231	0.125	0.082
		30	0.307	0.224	0.137	0.256	0.178	0.120	0.238	0.129	0.094
45	Gaussian	15	0.267	0.192	0.114	0.252	0.156	0.111	0.246	0.132	0.092
		20	0.241	0.172	0.115	0.239	0.151	0.094	0.232	0.122	0.085
		30	0.253	0.184	0.118	0.253	0.154	0.116	0.236	0.125	0.091
	Gamma	15	0.302	0.201	0.131	0.268	0.173	0.119	0.255	0.144	0.103
		20	0.298	0.203	0.113	0.257	0.162	0.112	0.254	0.131	0.097
		30	0.301	0.208	0.135	0.261	0.161	0.121	0.257	0.139	0.105
	EMG	15	0.241	0.182	0.098	0.224	0.140	0.098	0.221	0.118	0.090
		20	0.226	0.172	0.088	0.223	0.119	0.082	0.209	0.102	0.078
		30	0.256	0.183	0.101	0.216	0.145	0.092	0.221	0.122	0.089
50	Gaussian	15	0.222	0.154	0.069	0.216	0.112	0.066	0.211	0.098	0.064
		20	0.197	0.131	0.056	0.193	0.116	0.067	0.189	0.095	0.053
		30	0.206	0.145	0.065	0.205	0.116	0.071	0.192	0.098	0.056
	Gamma	15	0.255	0.152	0.087	0.218	0.124	0.079	0.210	0.112	0.072
		20	0.251	0.158	0.078	0.211	0.121	0.069	0.213	0.105	0.070
		30	0.256	0.164	0.090	0.215	0.128	0.083	0.214	0.102	0.072
	EMG	15	0.184	0.128	0.062	0.178	0.107	0.059	0.172	0.080	0.057
		20	0.177	0.116	0.066	0.172	0.083	0.055	0.158	0.077	0.037

SNR_1 (dB)	Shape	Rate	RMSE (Original) SNR_2 (dB)			RMSE (BNM) SNR_2 (dB)			RMSE (BNMD) SNR_2 (dB)		
			20	25	30	20	25	30	20	25	30
		30	0.197	0.127	0.068	0.177	0.103	0.055	0.169	0.081	0.051

TABLE 3

Percentage of the Number of Ions with Statistically Significant Variation in QCs for Synthetic Data

SNR_1 (dB)	SNR_2 (dB)	NISV (%)						
		Original	TIC	MedScale	Quantile	2D-GPRM-EIC	BNM	BNMD
40	20	17.1	15.7	14.5	13.2	12.1	10.8	9.9
	25	12.2	11.0	10.1	9.3	8.7	7.6	7.0
	30	8.9	8.2	7.5	7.0	6.4	5.9	5.3
45	20	14.3	12.8	11.9	10.2	9.2	8.0	6.8
	25	10.3	9.9	9.4	8.6	7.3	5.7	4.7
	30	8.5	7.9	6.8	6.0	5.4	4.4	4.0
50	20	12.2	11.4	10.2	9.1	8.9	7.8	6.5
	25	9.7	9.3	9.0	8.2	6.7	5.0	4.3
	30	8.0	7.4	6.9	5.8	5.1	4.1	3.8

TABLE 4

RMSE of Internal Standards (IS) for Metabolomic Data Before (Raw) and After Normalization by BNM and BNMD

	RMSE (Raw)					RMSE (BNM)					RMSE (BNMD)				
Shape	**IS_1**	**IS_2**	**IS_3**	**IS_4**	**IS_5**	**IS_1**	**IS_2**	**IS_3**	**IS_4**	**IS_5**	**IS_1**	**IS_2**	**IS_3**	**IS_4**	**IS_5**
Gaussian	0.129	0.138	0.107	0.152	0.114	0.103	0.119	0.091	0.128	0.101	0.071	0.085	0.068	0.092	0.073
Gamma	0.134	0.142	0.128	0.165	0.123	0.116	0.124	0.110	0.135	0.112	0.094	0.097	0.089	0.108	0.084
EMG	0.118	0.125	0.103	0.148	0.107	0.093	0.105	0.081	0.112	0.096	0.062	0.074	0.050	0.084	0.068

TABLE 5

Percentage of the Number of Ions with Statistically Significant Variation and Percentage of Decrease in MSD of QCs for Metabolomic Data

Measure	Batch	Raw	TIC	MedScale	Quantile	2D-GPRM-EIC	BNM	BNMD
NISV (%)	1 Positive	5.0	6.1	5.5	4.0	2.1	1.5	1.1
	2 Positive	5.0	3.1	4.1	2.9	2.2	1.7	1.2
	1 Negative	7.0	4.1	3.5	3.0	1.8	1.4	1.0
	2 Negative	20	11	9.9	7.3	4.4	3.2	2.6
ΔMSD (%)	1 Positive		7.8	11	13	21	24	29
	2 Positive		9.2	12	17	23	28	33
	1 Negative		7.3	11	14	21	26	30
	2 Negative		16	18	22	27	36	45

TABLE 6

Estimated EMG Peak Shape Parameters for the Metabolomic Data

Batch	Parameter	Average	β_{mz}	β_{RT}	ρ_α
1 Positive	μ	4.11 (0.15)	0.425 (0.044)	0.053 (0.003)	$\begin{bmatrix} 1.000 & +0.462 & -0.353 & -0.180 \\ & 1.000 & +0.160 & +0.084 \\ & & 1.000 & +0.058 \\ & & & 1.000 \end{bmatrix}$
	σ	1.67 (0.06)	−0.065 (0.008)	0.273 (0.028)	$\begin{bmatrix} 1.000 & +0.625 & +0.237 & +0.112 \\ & 1.000 & +0.342 & +0.205 \\ & & 1.000 & +0.193 \\ & & & 1.000 \end{bmatrix}$
	ζ	1.91 (0.08)	0.519 (0.057)	0.190 (0.013)	$\begin{bmatrix} 1.000 & +0.370 & -0.290 & -0.101 \\ & 1.000 & +0.530 & +0.119 \\ & & 1.000 & +0.143 \\ & & & 1.000 \end{bmatrix}$
2 Positive	μ	3.97 (0.14)	0.412 (0.046)	0.056 (0.004)	$\begin{bmatrix} 1.000 & +0.458 & -0.349 & -0.183 \\ & 1.000 & +0.157 & +0.078 \\ & & 1.000 & +0.062 \\ & & & 1.000 \end{bmatrix}$
	σ	1.69 (0.06)	−0.072 (0.006)	0.258 (0.029)	$\begin{bmatrix} 1.000 & +0.606 & +0.242 & +0.117 \\ & 1.000 & +0.337 & +0.192 \\ & & 1.000 & +0.201 \\ & & & 1.000 \end{bmatrix}$
	ζ	1.94 (0.07)	0.497 (0.036)	0.212 (0.017)	$\begin{bmatrix} 1.000 & +0.365 & -0.280 & -0.095 \\ & 1.000 & +0.542 & +0.124 \\ & & 1.000 & +0.136 \\ & & & 1.000 \end{bmatrix}$

Batch	Parameter	Average	β_{mz}	β_{RT}	ρ_α
1 Negative	μ	4.33 (0.13)	0.381 (0.052)	0.067 (0.008)	$\begin{bmatrix} 1.000 & +0.553 & -0.396 & +0.230 \\ & 1.000 & +0.162 & +0.064 \\ & & 1.000 & +0.071 \\ & & & 1.000 \end{bmatrix}$
	σ	1.59 (0.05)	−0.058 (0.002)	0.319 (0.030)	$\begin{bmatrix} 1.000 & +0.598 & +0.249 & +0.103 \\ & 1.000 & +0.325 & +0.181 \\ & & 1.000 & +0.188 \\ & & & 1.000 \end{bmatrix}$
	ζ	1.87 (0.07)	0.483 (0.046)	0.256 (0.020)	$\begin{bmatrix} 1.000 & +0.344 & -0.302 & -0.115 \\ & 1.000 & +0.497 & +0.132 \\ & & 1.000 & +0.122 \\ & & & 1.000 \end{bmatrix}$
2 Negative	μ	4.29 (0.15)	0.377 (0.050)	0.059 (0.006)	$\begin{bmatrix} 1.000 & +0.560 & -0.403 & -0.224 \\ & 1.000 & +0.155 & +0.059 \\ & & 1.000 & +0.073 \\ & & & 1.000 \end{bmatrix}$
	σ	1.61 (0.05)	−0.065 (0.003)	0.324 (0.031)	$\begin{bmatrix} 1.000 & +0.672 & +0.255 & +0.099 \\ & 1.000 & +0.308 & +0.171 \\ & & 1.000 & +0.166 \\ & & & 1.000 \end{bmatrix}$
	ζ	1.89 (0.06)	0.471 (0.048)	0.243 (0.017)	$\begin{bmatrix} 1.000 & +0.382 & -0.410 & -0.195 \\ & 1.000 & +0.426 & +0.144 \\ & & 1.000 & +0.155 \\ & & & 1.000 \end{bmatrix}$

TABLE 7

Mean CV and Percentage of Decrease in MSD for Proteomic Data for Different Normalization Methods

Method	Mean CV % (± S.D.)	ΔMSD (%)
Raw	47.9 (±14.9)	
TIC	44.6 (±11.8)	5.71
MedScale	39.5 (±10.2)	16.4
Quantile	36.3 (±8.65)	22.7
2D-GPRM-EIC	33.1 (±7.07)	28.9
BNM	29.0 (±5.50)	37.4
BNMD	26.1 (±4.93)	44.2