Analysis of LC-MS Data Using Probabilistic-Based Mixture Regression Models

Habtom W. Ressom, Getachew K. Befekadu, Mahlet G. Tadesse, Georgetown University, Washington DC

Dedicated to Prof. Dr.-Ing. habil. Lothar Litz on his 60th birthday

Summary A novel framework of a probabilistic-based mixture regression model (PMRM) is presented for alignment of multiple liquid chromatography-mass spectrometry (LC-MS) data with respect to retention time (RT) and mass-to-charge ratio (m/z). The expectation maximization algorithm is used to estimate the joint parameters of spline-based mixture regression models and prior transformation density models. The applicability of PMRM for alignment of LC-MS data is illustrated through three datasets. The performance of our method is compared with other approaches including dynamic time warping, correlation optimized warping, and continuous profile model in terms of coefficient variation of replicate LC-MS data and accuracy in detecting differentially abundant peptides/proteins.

Keywords Expectation-maximization, liquid chromatography, mass spectrometer, mixed-regression model

Schlagwörter Erwartungswert-Maximierungsalgorithmus, Flüssigchromatographie, Massenspektrometrie, Regressionsmodelle

1 Introduction
In proteomic studies, liquid chromatography coupled with mass spectrometry (LC-MS) is a common platform to identify and determine the abundance of various peptides that characterize particular proteins in biological samples [1; 2]. A mass spectrometry (MS) takes a sample as input and produces a measure of abundance of molecules that have particular mass-to-charge (m/z) values. In mass spectrometry-based proteomics, the molecules in question are peptides (short amino acid sequences), or sometimes whole proteins. For protein mixtures that are very complex, such as serum, a separation step is often used to physically separate parts of the sample before injection into the mass spectrometry. This separation, often conducted by a liquid chromatography (LC), means that a less complex mixture is fed into the mass spectrometry at any one time. The LC step spreads out parts of the sample solution over time (also called retention time) on the basis of some properties of the molecules. For example, in reverse phase high per-
performance liquid chromatography (RP-HPLC), a tubular column is packed with some material made up of hydrophobic molecules. The sample is then loaded onto the packed column, and sample molecules bind to the packed material with different affinities depending on each molecule’s hydrophobic properties. The solution is then pumped through the column. Initially, the solution is mostly water, and is increasingly made to contain more and more organic solvent. As this organic solvent gradient is applied, molecules that are more weakly bound to the packed material come off first, and progressively those that are more and more strongly bound come off. As liquid comes off the column, it accumulates into a droplet, which is then vaporized and ionized as it is introduced into the mass spectrometer. The advantage of using LC-MS is that at each RT point a less complex mixture (less complex than if we had injected the whole sample at once into the MS) is used to obtain mass spectral measurements. Peptides with the same \( m/z \) values are less likely to analyzed by MS at the same time, thus reducing potential ambiguity is needed. Also, the fewer the number of ions being simultaneously analyzed by MS at a time, the less ion suppression comes into play.

In a typical LC-MS based experiment, first mixtures of proteins are isolated from biological samples and then digested into peptides using enzymes. Following this, the peptides are separated by one or more steps of high-pressure LC, and are eluted into an electro-spray ion source, where they are nebulized in small and highly charged droplets. The output of an LC-MS experiment consists of three dimensions: (1) the elution time, also called retention (RT) point, (2) the \( m/z \) value, and (3) the intensity (ion abundance). Figure 1a presents three-dimensional data derived from a typical LC-MS experiment for a single run. As shown in the figure, each LC-MS run generates spectra comprised of thousands of peak intensities for peptides with specific RT and \( m/z \) values. Figure 1b shows a mass spectrum (ion abundance vs. \( m/z \)) at a particular RT point (RT in the figure is 10 minutes). Figure 1c depicts the total ion count (TIC) obtained by calculating the sum of the ion abundances across the \( m/z \) dimension for each RT point. Although RT is a continuous variable, the LC-MS system produces mass spectra at a discrete set of RT points, usually a few seconds apart. It is typical to represent RT points by scan indices, since there is a one-to-one correspondence between RT points and total MS scan numbers.

In differential protein expression studies, multiple LC-MS runs are compared to identify differentially abundant peptides between distinct biological groups. This is a challenging task because of the following reasons: (1) substantial variation in RT across multiple runs due to the LC instrument conditions and the variable complexity of peptide mixtures, (2) variation in \( m/z \) values due to occasional drift in the calibration of the mass spectrometry instrument, and (3) variation in peak intensities due to spray conditions (in most cases this is proportional to concentration of peptides in the sample). Thus, efficient and robust alignment algorithms are needed for quantitative comparison of multiple LC-MS runs. For example, Fig. 2a presents a heatmap of TIC profiles representing multiple samples. As illustrated in the figure, there is substantial drift in the RT dimension that needs to be aligned prior to selecting differently abundant ions. Figure 2b depicts a heatmap of the TIC profiles after alignment.

Various methods for alignment of LC-MS data have been described in literature including dynamic time warping (DTW) [3], correlation optimized warping (COW) [3], vectorized peaks [4], (semi-) supervised alignment using nonlinear regression methods [5], statistical alignment [6], and clustering [7]. Most of these algorithms are either limited to a consensus pair-wise combination of LC-MS runs or use a reference (template) for alignment. These limitations may lead to sub-optimal results compared to global alignment techniques. Methods that rely on optimization of global fitting functions provide an alternative solution to alignment of multiple LC-MS runs representing distinct biological groups. For example, a recently introduced method called continuous profile model (CPM) has been applied for alignment of continuous time-series data and for detection of differences in multiple LC-MS data [8; 9]. Although CPM is described as a naïve and computationally intensive method, the method has notable limitations. For example, it creates superfluous signal gaps, leading to non-uniform trace points across multiple LC-MS runs. Also, it is susceptible to fall into local minimum solutions due to the sub-optimal problem formulation. Another limitation of CPM algorithm is its poor performance with time complexity scales, requiring substantial computation time in modeling high resolution data. Thus, CPM is more suitable for low resolution of LC-MS data generated from less complex fractionations. Recently, Morris et al. developed a Bayesian method for analysis of matrix-assisted laser desorption ionization–time of flight (MALDI-TOF) proteomic data [10]. Their motivation extends from earlier work on Bayesian implementation of the wavelet-based functional mixed effects model introduced by Morris and Carroll [11]. The approach is similar to the spline-based functional mixed effects model introduced by Guo [12], which involves generalized mixed model equations to handle potentially irregular data. The method specifically deals with the identification of differentially expressed spectral regions across different experimental conditions assuming the alignment issue has already been taken care of.

In this paper, we propose a probabilistic-based mixture regression model (PMRM) for alignment of LC-MS data. Important features of this framework include the following: (1) it is not confined to landmark/feature alignment, (2) it allows continuous time alignment, (3) it utilizes a probabilistic-based functional modeling to deal with potentially irregular data, and (4) it is amenable to model both low and high resolution mass spectra, since it does...
not introduce any superfluous signal gap across multiple LC-MS runs. Moreover, the framework lends itself to an expectation-maximization (EM) algorithm with the following features: (i) the explicit use of transformation priors for modeling the ion abundance (peak intensity) variability in both RT and $m/z$ dimensions of the data, (ii) the use of a probabilistic metric that allows estimation of the distance among multiple spectra instead of computing pair-wise distances, and (iii) the ability to extend the method for alignment of LC-MS data involving...
Anwendungen

Anwendungen

Figure 2 Heat maps of TIC profiles before (a) and after alignment (b).

Figure 2 Heat maps of TIC profiles before (a) and after alignment (b).

Figure 2 Heat maps of TIC profiles before (a) and after alignment (b).

Figure 2 Heat maps of TIC profiles before (a) and after alignment (b).

Figure 2 Heat maps of TIC profiles before (a) and after alignment (b).

Figure 2 Heat maps of TIC profiles before (a) and after alignment (b).

multiple groups. In addition to the need for alignment of LC-MS data, one may also need to correct for systematic differences in the amplitude of the signal (ion abundances) at each RT point and m/z value which is called the problem of normalization for multiple LC-MS data. Correcting such systematic variations in the amplitude of LC-MS data at each RT point and m/z value is not the focus of this paper, although the two can be intertwined, i.e., one may perform alignment after normalization or the other way. Such normalization strategy could be addressed using either statistical models [13; 14] which were originally developed for gene expression data analysis or by using multiple internal standard compounds based on some empirical rules [15].

The remainder of the paper is organized as follows. In Sect. 2, we outline the probabilistic-based mixture regression model for alignment of LC-MS data. This section also explains the estimation algorithm for finding the maximum likelihood parameters of the spline-based mixture regression models and the prior transformation density models used for modeling the variability in RT, m/z, and measurement space (peak intensities of peptides) of the LC-MS data. Section 3 briefly describes the datasets used in this paper. Results and discussion illustrating the applicability of our approach for analysis of multiple LC-MS data are given in Sect. 4. We finally summarize our findings in Sect. 5.

2 Methods

2.1 Probabilistic-Based Mixture Regression Models

We propose a novel probabilistic-based mixture regression model for analysis of LC-MS data. A particular advantage of this approach is its ability to model non-Gaussian multi-modal density functions using simpler component density functions that can be defined on non-vector data such as LC-MS data. We assume that the observed dataset $D$ representing multiple groups is generated with the following features:

- An individual is randomly drawn from a population of $M$ objects (i.e., the dataset $D$).
The individual is assigned to the \(k\)-th group with probability \(a_k\), where \(\sum_{k=1}^{K} a_k = 1\). These are the prior weights corresponding to all \(K\) groups, where \(K < M\).

For an individual \(i\) where \(\{i | i \in k\)-th group\} there is a density function \(p_k(y_i | \theta_k)\) that generates the observed functional data \(y_i\).

From the above, it follows that the observed density on the \(y_i\)'s is a mixture model, i.e., a convex combination of component models \(p(y_i | \theta) = \sum_{k=1}^{K} a_k p_k(y_i | \theta_k)\). Thus, we estimate the most likely values for the parameters \(\theta_k\) and \(a_k\) using the assumed functional densities \(p_k(.)\) on the observed data \(y_i\)'s. This is accomplished by using the EM algorithm which is a general procedure for finding the maximum-likelihood estimation of the parameters from the mixture models [16–18]. Thus, this probabilistic based framework allows us to find the best group alignment in time and measurement spaces from the observed dataset \(D\).

In the following, we consider \(M\) LC-MS runs \(Y = \{y_i\}_{i=1}^{M}\) (i.e., observed dataset \(D\)), where the \(i\)-th run \(y_i\) has a length of \(n_i\), corresponding to the observation points (or time points) \(x_i\). To define the mixture regression model, each observation \(y_i\) is expressed as a function of some known \(x_i\), i.e., each component model is represented by a conditional density function of the form 

\[ p_k(y_i | x_i, \theta_k) \]

We assume a standard functional regression model between \(y_i\) and \(x_i\) as:

\[ y_i = \Psi_k(x_i, \beta_k) + e_i \]  

(1)

where \(e_i\) is a zero-mean Gaussian with variance \(\sigma^2\), i.e., \(e_i \sim N(0, \sigma^2)\) and \(\Psi_k(., .)\)'s are mapping functions of \(x_i\). Here, \(\theta_k\) includes both the parameters of the mapping model \(\Psi_k(x_i, \beta_k)\), i.e., the regression coefficients \(\beta_k\) and the noise variances \(\sigma^2\). Thus, the probability of observing \(y_i\), given \(x_i\) and component model \(k\), is assumed to be a conditional regression model. We define the conditional density of the observed data \(p_k(y_i | x_i, \theta_k)\) as a mixture density conditioned on the regression model parameters as well as on \(x_i\) values:

\[ p(y_i | x_i, \theta) = \sum_{k=1}^{K} a_k p_k(y_i | x_i, \theta_k) \]  

(2)

where \(p_k(y_i | x_i, \theta_k)\)'s are the mixing components, \(a_k\)'s are the mixing weights, \(\theta_k\) is the parameter set for the component \(k\), and \(\theta = \{\theta_k\}_{k=1}^{K} = \{\beta_k, \sigma^2\}_{k=1}^{K}\) is the complete (global) parameter set for the mixture model in Eq. (2).

Conditional independence between the functional data representing multiple groups allows the full joint density to be written as:

\[ p(Y | X, \theta) = \prod_{i=1}^{M} \sum_{k=1}^{K} a_k p_k(y_i | x_i, \theta_k) \]  

(3)

Thus, the log-likelihood of \(\theta\) for the given dataset \(D\) can be obtained directly from Eq. (3) as follows:

\[ \ell(\theta | D) = \log p(Y | X, \theta) \]

\[ = \sum_{i=1}^{M} \log \sum_{k=1}^{K} a_k p_k(y_i | x_i, \theta_k) \]

(4)

2.2 Expectation-Maximization Estimation for the Mixture Regression Models

The goal is to pull out the mixture of components from the full joint density, using the observed dataset \(D\) as a guide, so that the underlying group behavior can be inferred. A standard approach to deal with the hidden data is to utilize the EM algorithm for consistent estimation. Introducing hidden data, which correspond to the unknown group membership of the \(M\) functional data, is more appropriate in the log-likelihood formulation of Eq. (4). With this, let \(Z\) be a matrix of reporting vectors \(z_i = (z_{i1}, z_{i2}, \ldots, z_{iK})\) such that \(z_{ik} = 1\) for some \(k\) and \(z_{ik} = 0\), \(\forall i \neq k\). If \(z_{ik} = 1\), then the \(i\)-th spectrum is generated from the \(k\)-th mixture component. Thus, the joint density of \(Y\) and \(Z\) given \(X\) can be rewritten as follows:

\[ p(Y, Z | X, \theta) = \prod_{i=1}^{M} p(y_i | z_i, x_i, \theta) \times p(z_i) \]  

(5)

This follows from the conditional independence assumption on \(Y\) and \(z_i\)'s. Thus, the complete-data log-likelihood follows from Eq. (5) as:

\[ \ell(\theta | D, Z) = \log p(Y, Z | X, \theta) \]

\[ = \sum_{i=1}^{M} \sum_{k=1}^{K} z_{ik} \log \alpha_k \]

\[ + \sum_{i=1}^{M} \sum_{k=1}^{K} z_{ik} \log p(y_i | x_i, \theta_k) \]

(6)

To fit the dataset \(D\) into this framework for the spline-based mixture regression model, i.e., the regression model between \(y_i\) and \(x_i\) defined in Eq. (1), is rewritten as follows:

\[ y_i = B_k \beta_k + e_i \quad e_i \sim N(0, \sigma^2) \]

(7)

where \(B_k\) denotes spline basis matrix evaluated at \(x_i\). Then, the group specific conditional probability density function for \(y_i\) based on the error model can be rewritten as follows:

\[ p_k(y_i | x_i, \theta_k) = N(y_i | B_k \beta_k, \sigma^2) \]

(8)

Thus, Eq. (8) defines the probabilistic model for spline-based mixture regression. Here we remark that the estimation algorithm is implemented by taking advantage of the connection between smoothing B-splines (at the design points) and mixture regression models. Splines are recommended for data fitting whenever there is no particular reason for using a single polynomial or other elementary functions. Spline functions have the following useful properties: smooth and flexible, easy to evaluate, along with their derivatives and integrals, and easy to generalize to higher dimensions.

2.3 Modeling Variability Along the RT Dimension

In the previous sections, we introduce a probabilistic-based framework that allows us to incorporate arbitrary scaling and translation variability in time and measurement spaces of the LC-MS data representing single or
multiple groups. This framework can be used to align multiple chromatograms involving two dimensions (TIC values vs. RT points). In this section, we describe the steps that define the principle upon which the alignment models are based for the LC-MS data. These steps provide the data generative model, define the prior transformation models, compute the joint probabilities and the log-likelihood function, and describe the associated EM algorithm.

With a minor abuse of notation, Eq. (9) presents the general model formulation and the assumed prior transformation models for scaling and translation parameters:

\[ y_i = d_i \cdot 1 + c_i \Psi_k((a_i x_i - b_i) \cdot 1) + e_i \]

(9)

where \( a_i, b_i \) and \( c_i, d_i \) are scaling and translation variables for time and measurement spaces, respectively; and \( 1 \) is a column vector of ones with an appropriate dimension. The posterior probabilities of these parameters are estimated together with the other regression model parameters as well as the error variances from the observed dataset \( D \).

The definition given in Eq. (9) and the corresponding mixture regression model allow arbitrary scaling and translation in time and measurement spaces. Here we remark that the conditional probabilistic density model \( p(y_i | x_i, \theta_k, a_i, b_i, c_i, d_i) \) is also conditioned on the global parameter sets \( \theta_k = \{ \beta_k, \sigma^2_k \} \) for \( k = 1, 2, \ldots, K \). The transformation priors \( a_i, b_i \) and \( c_i, d_i \) are associated to each LC-MS run using the model of Eq. (9). Thus, together with prior transformation models, the problem of deriving the resulting joint probabilities can be directly handled. By considering \( a_i, b_i \) and \( c_i, d_i \) as random variables with their associated prior probability density functions, one can then proceed with the mixture regression models for alignment of data representing multiple groups. With the assumption of independence, the joint prior probabilistic model for the time scaling and translation \( p(a_i, b_i) = p(a_i) \times p(b_i) \) should encode the idea that the most likely translation is the zero-time translation and should also discount large translations. A zero-mean Gaussian prior is a good fit for this, i.e., \( b_i \sim N(0, \sigma^2_b) \) where \( \sigma^2_b \) is the variance. Moreover, for the time scaling prior \( a_i \), a value of one is the most likely value (i.e., no scaling at all). Thus, the corresponding Gaussian prior model is given as \( a_i \sim N(1, \sigma^2_a) \cdot I(a_i > 0) \). The variances \( \sigma^2_a \) and \( \sigma^2_b \) will be learned from the data within the ensuing EM algorithm.

Similarly, the joint prior probabilistic model for the affine transformation in measurement space with independence assumption is given as \( p(c_i, d_i) = p(c_i) \times p(d_i) \). Here, we follow a similar argument as that of the affine time transformation above. In other words, we assume that \( c_i \sim N(0, \sigma^2_c) \cdot I(c_i > 0) \) and \( d_i \sim N(0, \sigma^2_d) \) are the prior probabilistic models and specify the most likely transformations. We also remark that affixing \( k \) to each parameter, i.e., \( \theta_k = \{ \beta_k, \sigma^2_k, \tau^2_k, \omega^2_k, v^2_k \} \), is used to arrive at group-dependent regression model. After obtaining both the group-specific conditional probability density function for \( y_i \), i.e., \( p_i(y_i | x_i, \theta_k, a_i, b_i, c_i, d_i) \), and the complete data likelihood function \( \ell(\theta) \) (where \( \theta = \{ \beta_k, \sigma^2_k, \tau^2_k, \omega^2_k, v^2_k \} \) is the complete parameter set for the model in Eq. (9)), the EM algorithm sequentially executes the following two steps:

**Expectation-Step (E-Step):** take the expectation of \( \ell(\theta|D, Z) \) with respect to \( p(Z|Y, X, \theta^{-1}) \), where \( \theta^{-1} \) is the current set of parameters, and

**Maximization-Step (M-Step):** maximize the expected value of the computed log-likelihood from E-Step over the parameters \( \theta \) to yield new parameters \( \theta' \).

For the particular form of \( Z \) chosen here, the expectation of \( \ell(\theta|D, Z) \) is

\[
E[\ell(\theta|D, Z)] = \sum_{i=1}^{M} \sum_{k=1}^{K} h_{ik} \log \alpha_k + \sum_{i=1}^{M} \sum_{k=1}^{K} \sum_{l=1}^{K} \alpha_{lk} h_{il} \log p(y_i | x_i, \theta_k)
\]

(10)

where \( h_{ik} = E[z_{ik}] \propto p_i(y_i | x_i, \theta_k) \times p(z_{ik} = 1) \) and it corresponds to the posterior probability of the \( i \)-th spectra being generated from the component model \( k \). Therefore, the E- and M-steps of the EM estimation algorithm proceed iteratively to estimate the parameters \( \theta \). The following are the steps involved in estimating the parameters for the spline-based mixture regression model using the EM algorithm:

**Step 1:** Initialize randomly the membership probabilities \( h_{ik} = E[z_{ik}] \) and estimate the parameter set \( \theta' \) using these particular membership probabilities.

**Step 2:** Compute the new membership probabilities \( h_{ik} = E[z_{ik}] \) using the current parameter set \( \theta^{-1} \).

**Step 3:** Estimate the new parameter set \( \theta' \) using the new membership probabilities from Step 2.

**Step 4:** Repeat Steps 2 and 3 until a stopping criterion is met (e.g., maximum number of iteration or convergence rate).

### 2.4 Alignment of Three Dimensional LC-MS Spectra

Thus far, we have implicitly assumed that \( y_i \) consists of two-dimensional data (i.e., the model involves RT points and TIC values). To take advantage of the entire LC-MS run for differential protein expression studies, we extend our method for three-dimensional alignment (i.e., aligning with respect to RT points and \( m/z \) bins; and performing implicitly normalization of peak intensities for peptides). We denote a three-dimensional LC-MS run as \( \tilde{y}_i \) which consists of \( Q \) columns such that the \( q \)-th column \( \tilde{y}^{(q)}_i \) contains two-dimensional data for the \( i \)-th observation. In other words, each \( \tilde{y}^{(q)}_i \) corresponds to a standard two-dimensional profile stacked into a vector.

The three-dimensional nature of each LC-MS run can be incorporated into the alignment model by defining a separate regression model for each dimension. Thus, the general regression model based on the assumed priors for
scaling and translation parameters for each of the three dimensions (RT, m/z and peak intensities of peptides) can be presented as follows:

\[
\tilde{y}_{i} = d_{i}^{(q)} \cdot \mathbf{1} + c_{i}^{(q)} \psi_{k}^{(q)} \left( a_{i} x_{i} - b_{i} \cdot 1 \right), \tilde{y}_{k}^{(q)} + e_{i}^{(q)}.
\]

\[a_{i} \sim N \left( 1, \tau_{x}^{2} \right), \quad b_{i} \sim N \left( 0, \tau_{b}^{2} \right), \quad c_{i}^{(q)} \sim N \left( 1, \sigma_{k}^{(q)} \cdot \mathbf{1} \right), \quad d_{i}^{(q)} \sim N \left( 0, \sigma_{k}^{(q)} \cdot \mathbf{1} \right), \quad e_{i}^{(q)} \sim N \left( 0, \sigma_{k}^{(q)} \cdot \mathbf{1} \right), \]

\[i = 1, 2, ..., M, \quad q = 1, 2, ..., Q, \quad \text{and} \quad k = 1, 2, ..., K \]

where \( \beta_{i}^{(q)} \) consists of the regression coefficients for the \( q \)-th dimension (i.e., the regression coefficients for the \( q \)-th column of \( \psi_{k}^{(q)} \)); \( a_{i}, b_{i} \) and \( c_{i}^{(q)}, d_{i}^{(q)} \) are group specific scaling and translation variables for RT points, \( m/z \) bins and measurement space (peak intensities of peptides), respectively. Note that the scaling and translation parameters for measurement space vary from dimension to dimension.

The mixture of directed acyclic graph (DAG) model that shows the graphical factorization of the joint probability distribution for \( \tilde{y}_{i} \), is shown in Fig. 3. This graphical model consists of two components: the structure and the set of local distribution families. The structure represents conditional probabilities that assert through a factorization of the joint distribution for \( \tilde{y}_{i} \), while the local distribution families associated with mixture model are those in Eq. (11). In the following, we illustrate the derivation using only the translation in measurement space, \( \{d_{i}^{(q)}\}_{q=1}^{Q} \), for three-dimensional LC-MS data \( \tilde{y}_{i} \), using the model specified in Eq. (11):

\[
p(\tilde{y}_{i}, d_{i}^{(1)}, ..., d_{i}^{(Q)}) = \prod_{q=1}^{Q} N(\psi_{i}^{(q)} | B_{i}^{(q)} \beta_{k}^{(q)}, d_{i}^{(q)} \cdot \mathbf{1}, \sigma_{k}^{(q)} \cdot \mathbf{1}) \times N(d_{i}^{(q)} | 0, \sigma_{k}^{(q)} \cdot \mathbf{1})
\]

The joint density can be factored due to the following necessary conditions: (i) conditional independence is assumed between the dimensions of \( \tilde{y}_{i} \), and (ii) each dimension is assumed to have its own set of translation parameters. Thus, the marginal density \( p(\tilde{y}_{i}) \) is given as \( \prod_{q=1}^{Q} p(\tilde{y}_{i}^{(q)}) \) and the resulting log-likelihood of \( Y = \{\tilde{y}_{i}\}_{i=1}^{M} \) takes the following form

\[
\log p(Y) = \sum_{i=1}^{M} \log p(\tilde{y}_{i}) = \sum_{i=1}^{M} \sum_{q=1}^{Q} \int_{d_{i}^{(q)}} \log p(\tilde{y}_{i}^{(q)} | d_{i}^{(q)}) \times p(d_{i}^{(q)}) \, dd_{i}^{(q)}
\]

Equation (13) includes all the \( M \times Q \) spectra in the log-likelihood function and analytically integrating Eq. (13) gives the log-likelihood as:

\[
\log p(Y) = \sum_{i=1}^{M} \sum_{q=1}^{Q} \log N(\psi_{i}^{(q)} | B_{i}^{(q)} \beta_{k}^{(q)}, \psi_{i}^{(q)} \cdot 1 + \sigma_{k}^{(q)} \cdot 1)
\]

Thus, for a complete alignment problem in the measurement space including both scaling and translation parameters, the individual dimensions require \( 2Q \) separate scaling and translation parameters to be incorporated in this framework.

On the other hand, the condition for RT point alignment needs a different perspective. Here the assumption is that the dynamic behavior for each dimension occurred over the same time scale. Thus, the RT transformation parameters need to be shared over the \( Q \) dimensions for each group data. For the RT translation case, each two-dimensional profile in \( \tilde{y}_{i} \) shares a single translation parameter \( b_{i} \) from its corresponding group behavior and the conditional density of \( \tilde{y}_{i} \) becomes

\[
p(\tilde{y}_{i} | b_{i}) = \prod_{q=1}^{Q} p(\tilde{y}_{i}^{(q)} | b_{i}) = \prod_{q=1}^{Q} N(\psi_{i}^{(q)} | B_{i}^{(q)} \beta_{k}^{(q)}, \psi_{i}^{(q)} \cdot 1 + \sigma_{k}^{(q)} \cdot 1)
\]

in which there is only one \( b_{i} \) for all \( q \) in each group. The corresponding marginal density \( p(\tilde{y}_{i}) \) cannot be factored as that of measurement space translation due to the dependence among dimensions through the translation parameters \( b_{i} \). Therefore the log-likelihood of \( Y \) is given as follows:

\[
\log p(Y) = \sum_{i=1}^{M} \log p(\tilde{y}_{i}) = \sum_{i=1}^{M} \log \int_{b_{i}} \left[ p(b_{i}) \prod_{q=1}^{Q} p(\tilde{y}_{i}^{(q)} | b_{i}) \right] \, db_{i}
\]

The product over the dimension is now inside the integration operation and an approximate log-likelihood can be computed using the Monte-Carlo numerical integration technique [19]. Thus, the approximation can be computed as follows:

\[
\log p(Y) = \sum_{i=1}^{M} \log \left[ \sum_{r=1}^{R} \prod_{q=1}^{Q} p(\tilde{y}_{i}^{(q)} | b_{i}^{(r)}) \right] - M \log R
\]
where \( b^{(r)}_i \sim N(0, s^2_i) \) for \( r = 1, 2, \ldots, R \), and \( R \) is the sample size in the Monte-Carlo numerical integration. Moreover, we used the following approximation for the unconditional marginal density of \( y^{(q)}_i \) in arriving at Eq. (17):

\[
p(y^{(q)}_i) = \frac{1}{R} \sum_{k=1}^{R} \alpha_k p_k \left( y^{(q)}_i | b^{(r)}_i \right)
\]

The derivation for the scaling \( c^{(q)}_i \) in measurement space and scaling \( a \) in RT point will follow directly from the previous derivations except for handling the various individual \( q \) subscripts on the dimension dependent parameters. Moreover, the EM algorithm outlined in Sect. 2.3 is then modified at each iteration to update the following parameter set \( \theta^{(t)} = \left\{ b_i^{(t)}, \sigma_i^{(t)}, r_i^{(t)}, s_i^{(t)}, u_i^{(t)}, v_i^{(t)} \right\} \) for \( q = 1, 2, \ldots, Q \) and \( k = 1, 2, \ldots, K \).

### 3. LC-MS Data

#### 3.1 Single-class Listgarten et al. LC-MS Dataset

The single-class Listgarten et al. LC-MS dataset was obtained from [http://www.cs.toronto.edu/~jenn/LCMS](http://www.cs.toronto.edu/~jenn/LCMS). They consist of 11 replicate LC-MS runs generated from proteins of *lysed E.coli* cells using a capillary-scale LC coupled on-line to an ion trap mass spectrometer. Each run is represented by a 400 × 2400 data matrix corresponding to 400 RT points (∼55 min.) and 2400 m/z bins spanning between 400 and 1600 Da. Detailed experimental information can be found in ref. [20].

#### 3.2 Two-class Listgarten et al. LC-MS Dataset

The two-class Listgarten et al. dataset (downloaded from [http://www.cs.toronto.edu/~jenn/LCMS](http://www.cs.toronto.edu/~jenn/LCMS)) was obtained from human serum samples that represent two classes. The dataset consists of 14 LC-MS runs generated from two groups of human serum samples. Seven runs were derived from a sample, in which the human serum was mixed with three known peptides: Angiotensin II, Angiotensin I, and Substance P with molecular weights of 1047.2, 1297.51, and 1348.66 Da, respectively. The remaining seven runs were from a sample consisting of human serum only. The LC-MS data were generated using high performance liquid chromatography (HPLC) pump interfaced with a Thermo Finnigan LCQ quadrupole ion trap tandem mass spectrometer. Each LC-MS run is represented by a 500 × 2400 data matrix corresponding to 500 RT points (∼55 min.) and 2400 m/z bins between 400 and 1600 Da.

#### 3.3 Mueller et al. LC-MS Dataset

The Mueller et al. LC-MS dataset was obtained from [http://prototools.ethz.ch/luellu/web/Latin_Square_Data.php](http://prototools.ethz.ch/luellu/web/Latin_Square_Data.php). They consist of 18 LC-MS runs generated from tryptic digests of six standard non-human proteins (myoglobin, carbonic anhydrase, cytochrome c, lysozyme, alcohol dehydrogenase, and aldolase A) spiked with different concentrations into a complex sample background of human peptides and isolated from serum using solid-phase N-glycopeptidation. The LC-MS data were obtained using the Fourier transformed-linear trap quadrupole (FT-LTQ) mass spectrometer. Detailed experimental information can be found in ref. [6]. The 18 LC-MS runs were grouped into six classes based on the concentration of the spike-in proteins. Specifically, Latin square dilution with different concentrations of the six standard proteins was used. Each run is represented by a 2000 × 750 data matrix corresponding to 2000 RT points (∼55 min.) and 750 m/z bins between 300 and 1600 Da.

### 4 Results and Discussion

We applied PMRM to align one of the above three datasets based on two-dimensional profiles involving TIC and RT points. The remaining two datasets were analyzed on the basis of three-dimensional data involving RT points, m/z bins, and peak intensities of peptides.

#### 4.1 Single-class Listgarten et al. LC-MS Dataset

The single-class Listgarten et al. dataset was converted into 11 two-dimensional profiles by calculating the TIC at each RT point (i.e., at each RT point the sum of all ions across m/z values was calculated). Figure 4 depicts TIC profiles of the 11 replicate LC-MS runs. From the figure we observe significant shifts along the experimental RT points as well as distortions in the measurement space (TIC values). The data aligned by the PMRM approach are shown in Fig. 5. Figure 6 depicts the mean TIC profiles for the original (before alignment) and those aligned by DTW, COW, CPM, and PMRM. From this figure, we see that DTW and CPM approaches produce superfluous signal gaps, thus stretching the experimental RT points by almost 1.5 and 2 folds, respectively. Although the stretched RT points can be rescaled to the original RT points, the difficulty arises in sorting out the inserted gaps during peptide sequence identification to accurately match the sequence information with the experimental RT points. On the other hand, the aligned tracing points obtained by PMRM and COW are consistent with the experimental RT points.

**Table 1** Coefficient of variations for single-class and two-class Listgarten et al. datasets before alignment (original) and after alignment by DTW, COW, CPM, and PMRM. The result of the method with the best performance in each category is shown in boldface.

<table>
<thead>
<tr>
<th></th>
<th>Single-class Listgarten et al.</th>
<th>Two-class Listgarten et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class 1</td>
<td>Class 2</td>
</tr>
<tr>
<td>Original</td>
<td>82%</td>
<td>57%</td>
</tr>
<tr>
<td>DTW</td>
<td>70%</td>
<td>56%</td>
</tr>
<tr>
<td>COW</td>
<td>80%</td>
<td>56%</td>
</tr>
<tr>
<td>CPM</td>
<td>57%</td>
<td>44%</td>
</tr>
<tr>
<td>PMRM</td>
<td>65%</td>
<td>47%</td>
</tr>
</tbody>
</table>
Table 1 presents the coefficient of variations (CVs) for the original data and those aligned by DTW, COW, CPM, and PMRM. As shown in the table, both CPM and PMRM have yielded better results than DTW and COW.

4.2 Two-class Listgarten et al. LC-MS Data

Alignment. Each LC-MS run in the two-class Listgarten et al. dataset was reduced from a $500 \times 2400$ to a $500 \times 8$ data matrix by partitioning the $m/z$ dimension into eight equal $m/z$ bins. This reduction was necessary to handle the computational complexity. We applied the PMRM to the reduced matrix and obtained a set of aligned data matrices, i.e., for each LC-MS run ($i = 1, 2, \ldots, 14$), a data matrix $M_i$ was obtained, where $M_{i, \tau, q}$ is the ion abundance after alignment at the $\tau$-th aligned RT point ($\tau = 1, 2, \ldots, 500$) and the $q$-th $m/z$ bin ($q = 1, 2, \ldots, 8$).
For visualization purpose, we calculated the TIC values from the reduced original and aligned matrices. Figures 7 and 8 depict the TIC profiles of the original and aligned LC-MS runs, respectively. Table 1 presents the CVs for the original and those aligned by DTW, COW, CPM, and PMRM. As illustrated in the table, both CPM and PMRM have yielded better results than DTW and COW. Although, CV is a good measure for assessing the variability among replicate datasets, knowledge of such information does not necessarily confirm or directly link with actual interpretation of differences between the LC-MS datasets. In the following, we use statistical measures for detected differences between the two classes to assess the performance of various alignment methods.

**Difference detection.** We used the aligned data matrices for difference detection. Specifically, we used a multivariate permutation method implemented in the BRB-Array-Tools (National Cancer Institute, Bethesda, MD) to calculate non-parametric p-values for each RT point and m/z bin in the aligned data matrices representing the two-class dataset. We utilized the experimentally extracted ground truth information to evaluate the sensitivity, specificity, and accuracy in detecting the truly differentially abundant peptides. This is accomplished by using a list of experimentally extracted ground truth peptide m/z values provided in ref. For evaluation, we considered three significance levels (p < 0.05, p < 0.01, and p < 0.001). For each significance level, we performed a hypothesis test to determine whether the ion abundances differ between the two-class datasets. Table 2 presents the computed sensitivity, specificity, accuracy, and precision in detecting differentially abundant peptides using the LC-MS data aligned by DTW, COW, CPM and PMRM. The table also provides an F-measure, which is a weighted harmonic mean measure combining the precision and recall, calculated as follows with $\beta = 1$ [21]:

$$F_{\text{measure}} = \frac{(\beta^2 + 1) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}$$

As shown in Table 2, the LC-MS data aligned by PMRM have yielded overall better performance than those...
Table 2  Sensitivity, specificity, accuracy, precision and F-measure in detecting differentially abundant peptides after aligning in the Two-class Listgarten et al. dataset by the DTW, COW, CPM, and PMRM (for $p < 0.05$, $p < 0.01$, and $p < 0.001$). The result of the method with the best performance in each category is shown in boldface.

<table>
<thead>
<tr>
<th>Method</th>
<th>P-values</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTW</td>
<td>0.05</td>
<td>60%</td>
<td>53%</td>
<td>58%</td>
<td>60%</td>
<td>60%</td>
</tr>
<tr>
<td>COW</td>
<td></td>
<td>64%</td>
<td>57%</td>
<td>60%</td>
<td>62%</td>
<td>63%</td>
</tr>
<tr>
<td>CPM</td>
<td></td>
<td>59%</td>
<td>43%</td>
<td>70%</td>
<td>72%</td>
<td>64%</td>
</tr>
<tr>
<td>PMRM</td>
<td></td>
<td>65%</td>
<td>79%</td>
<td>73%</td>
<td>73%</td>
<td>69%</td>
</tr>
<tr>
<td>DTW</td>
<td>0.01</td>
<td>63%</td>
<td>63%</td>
<td>64%</td>
<td>63%</td>
<td>69%</td>
</tr>
<tr>
<td>COW</td>
<td></td>
<td>63%</td>
<td>61%</td>
<td>62%</td>
<td>77%</td>
<td>69%</td>
</tr>
<tr>
<td>CPM</td>
<td></td>
<td>45%</td>
<td>90%</td>
<td>68%</td>
<td>73%</td>
<td>57%</td>
</tr>
<tr>
<td>PMRM</td>
<td></td>
<td>62%</td>
<td>81%</td>
<td>72%</td>
<td>73%</td>
<td>67%</td>
</tr>
<tr>
<td>DTW</td>
<td>0.001*</td>
<td>67%</td>
<td>66%</td>
<td>66%</td>
<td>63%</td>
<td>65%</td>
</tr>
<tr>
<td>CPM</td>
<td></td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>PMRM</td>
<td></td>
<td>59%</td>
<td>83%</td>
<td>71%</td>
<td>73%</td>
<td>64%</td>
</tr>
</tbody>
</table>

* The condition for $p < 0.001$ is too stringent for CPM.

aligned by DTW, COW and CPM. For $p < 0.001$ significant level, CPM did not yield any differentially abundance ions. Possible reasons for this could be due to over smoothing and attenuation of peaks by CPM.

4.3 Mueller et al. LC-MS Dataset

Alignment. The Mueller et al. LC-MS dataset was reduced from a $2000 \times 750$ to $2000 \times 8$ data matrix by partitioning the $m/z$ dimension into eight equal $m/z$ bins. After applying PMRM to this reduced matrix, we obtained a set of aligned data matrices, i.e., for each LC-MS run ($i = 1, 2, \ldots, 18$), a data matrix $M^i$ was obtained, where $M^i_{\tau,q}$ is the ion abundance obtained after alignment at $\tau$-th aligned RT points ($\tau = 1, 2, \ldots, 2000$) and the $q$-th $m/z$ bin ($q = 1, 2, \ldots, 8$). Figures 9 and 10 depict TIC profiles of the original and aligned LC-MS spectra, respectively, for RT points between 800 and 2000.

Difference detection. The aligned data matrices were used for difference detection. Since the original LC-MS dataset consisted of the diluted series of six proteins with different concentration in each sample, we determined the peptides with statistically significant difference in their expression level based on an F-statistic. To accomplish

Table 3  Sensitivity, specificity, accuracy, precision and F-measure in detecting differentially abundant peptides in the Mueller et al. dataset by the DTW, COW, CPM, and PMRM for $p < 0.01$. The result of the method with the best performance in each category is shown in boldface.

<table>
<thead>
<tr>
<th>Method</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTW</td>
<td>76%</td>
<td>70%</td>
<td>85%</td>
<td>70%</td>
<td>73%</td>
</tr>
<tr>
<td>COW</td>
<td>83%</td>
<td>75%</td>
<td>79%</td>
<td>76%</td>
<td>79%</td>
</tr>
<tr>
<td>CPM</td>
<td>90%</td>
<td>75%</td>
<td>83%</td>
<td>84%</td>
<td>80%</td>
</tr>
<tr>
<td>PMRM</td>
<td>91%</td>
<td>78%</td>
<td>85%</td>
<td>83%</td>
<td>87%</td>
</tr>
</tbody>
</table>
this, we first generated a list of peptides assuming no mis-cleavage from trypsin digestion of six non-human proteins (myoglobin, carbonic anhydrase, cytochrome c, lysozyme, alcohol dehydrogenase, and aldolase A) using the MS-Digest software from the University of California, San Francisco (http://prospector.ucsf.edu). Then, we used the permutation method to examine each RT point and \( m/z \) bin. We compared the statistically significantly different RT-\( m/z \) pairs (\( p < 0.01 \)) against the list of peptides that represent the six proteins to calculate sensitivity, specificity, accuracy, precision, and F-measure. Table 3 presents these results for the data aligned by DTW, COW, CPM and PMRM. The table illustrates that PMRM yields overall better results than DTW, COW, and CPM.

### 4.4 Reproducibility Study

We examine the reproducibility behavior of the PMRM in aligning LC-MS data. To accomplish this, it is necessary that the identifiability property holds with respect to the mixture models given by Eq. (2) (or Eq. (3)) and the general model definition given by Eq. (9)). However, the general statement on the identifiability property requires minimizing distributions using Kullback-Leibler divergence from the true mixing distribution [22; 23]. Since the identifiability property of mixing distributions is of interest in its own right, we focus here on examining the consistency of the PMRM parameters by running the algorithm multiple times for the same set of LC-MS data. Table 4 presents the 95\%, 99\% and 99.9\% confidence intervals for the mean of alignment scaling field (i.e., the expected posterior for scaling \( a_i \) parameters) obtained in aligning the single-class Listgarten et al. LC-MS data. From this table, we can see that estimated values show fairly narrow intervals, indicating consistent estimation of the parameters across multiple runs. This analysis also provides similar mean posterior estimates for the other alignment field parameters. Moreover, it took approximately nine minutes for the algorithm to complete one run on a PC with an Intel Core 2 Duo 64 bit 2.66 GHz CPU and 8 GB RAM.

### 5 Conclusion

This paper proposes a probabilistic-based mixture regression model (PMRM) for alignment of multiple LC-MS data. We utilize the well known maximum likelihood-based EM algorithm for estimating the mixture regression models and the prior transformation (scaling and translation parameters) models of the LC-MS data. The latter accounts for the variability in RT points, \( m/z \) values, and peak intensities. The proposed framework allows alignment with respect to RT points and \( m/z \) bins; and it implicitly performs normalization of peak intensities in multiple LC-MS runs. The performance of the approach is assessed through three LC-MS datasets: replicate LC-MS dataset generated from proteins of lysed E.coli cells, LC-MS dataset representing two classes of human serum samples (with and without spike-in peptides), and spectra representing six classes, where six proteins are spiked at different concentrations into a complex sample background of human peptides. Through these datasets, it is demonstrated that PMRM achieves good coefficient
of variation among replicate LC-MS data while preserving the original experimental retention time, i.e. without introducing any superfluous signal gap across multiple LC-MS spectra. Also, through spike-in peptides, the paper demonstrates that PMRM leads to more accurate identification of differentially abundant peptides than DTW, COW, and CPM. Future work will focus on extending PMRM for alignment and normalization of multiple LC-MS data that will account variability among LC-MS runs representing samples from distinct biological groups as well as modeling the heterogeneity within group.

Acknowledgements

This work was supported in part by the National Science Foundation Grant (IIS-0812246), National Cancer Institute (NCI) R21 CA130837 Grant, NCI Early Detection Research Network Associate Membership Grant, and the Prevent Cancer Foundation Grant awarded to HWR.

References
