Statistics and Its Interface Volume 11 (2018) 581–586

Addressing varying non-ignorable missing data mechanisms using a penalized EM algorithm: application to quantitative proteomics data

So Young Ryu*

In multi-laboratory collaborative or large-scale proteomic studies, it is challenging to analyze data properly due to varying non-ignorable missing data mechanisms across experiments. PEMM (Penalized EM algorithm incorporating missing data mechanism) proposed by Chen, Prentice and Wang [1] estimates both the mean and the covariance of protein abundances in the presence of non-ignorable missing data; however, PEMM assumes a common missing mechanism for all experiments. This approach may be adequate when experiments are performed under similar conditions, but it may not work optimally when experiments are conducted in different laboratories or over a long period of time. In this paper, we extend PEMM to appropriately handle varying missing data mechanisms for datasets generated at multiple laboratories. Recognizing that jointly estimating missing mechanisms and parameters of interest is a challenging task, we assume that missing data mechanisms are known, and demonstrate benefits of incorporating multiple missing mechanisms for datasets generated at different laboratories. We call our algorithm PEMvM (Penalized EM algorithm for varying non-ignorable missing mechanisms). Our extension is simple and enjoys all the properties that PEMM offers. When missing data mechanisms differ across experiments, PEMvM performs better than PEMM in terms of accurate mean estimation and data imputation. In this paper, we demonstrate the performance of PEMvM using both simulated and real proteomic data.

Keywords and phrases: Mass spectrometry, Proteomics, Protein relative quantitation, Non-ignorable missing data, Varying missing data mechanisms.

1. INTRODUCTION

Mass spectrometry (MS) can quantify thousands of proteins in complex biological samples and detect biomarker proteins. Driven by the big data movement in proteomics, several multi-laboratory consortia or large-scale proteomic datasets became available; however, it is challenging to analyze such data collectively because of a high rate of missing values and non-ignorable missing data patterns [6, 5]. To make it even harder for multi-laboratory collaborative studies, the missing data mechanisms across laboratories may vary due to different instrument settings or experimental procedures. Varying missing data mechanisms across laboratories may also be observed in large-scale studies due to several factors (i.e. different instrument conditions over time, column degradation) [10].

In mass spectrometry data, less abundant proteins have higher probabilities of being missing, thus the missing data mechanisms of protein abundances are not random (NMAR). Ignoring NMAR mechanisms may result in invalid statistical inferences [9, 4], and inaccurate scientific conclusions. Previously, several research groups proposed methods that incorporated non-ignorable missing data mechanisms. Luo et al. [6] proposed a Bayesian hierarchical model assuming a linear relationship between the peptide missing probability and the observed abundance at the logit scale. Ryu et al. [10] developed a censored regression model considering intensity-dependent missing values. Recently, Chen, Prentice and Wang [1] proposed a penalized EM algorithm incorporating a non-ignorable missing data mechanism (PEMM) to jointly estimate the protein mean abundance and its covariance matrix. PEMM constrained the parameter space of $\boldsymbol{\Sigma}$ and imposed a lower and upper bound of each eigenvalue using a penalty amounting to an inverse-Wishart prior. The performance of PEMM was demonstrated by comparing the mean and covariance estimates to various methods such as available case analysis, k-nearest neighbor imputation, and penalized EM algorithm with MAR assumption. The covariance matrix estimated by PEMM can be useful to study the association among proteins; however, to date, there is no method that incorporates different non-ignorable missing value mechanisms across laboratories. In addition, as noted in Chen, Prentice and Wang [1], it is challenging to jointly estimate missing mechanisms and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

In this paper, we extended the PEMM algorithm to handle multiple non-ignorable missing data mechanisms assuming that such missing mechanisms were known. We first outlined the PEMM algorithm and then introduced our extension, PEMvM (Penalized EM algorithm for varying non-ignorable missing mechanisms). Then, we compared the performances of these two algorithms in mean/covariance esti-

*Corresponding author.

mations and missing data imputations. We used both simulated data and inter-laboratory data from the Clinical Proteomic Technologies Cancer Consortium [1, 8] to demonstrate the performance of PEMvM.

2. METHODS

2.1 PEMM

The following is the PEMM algorithm developed by Chen, Prentice and Wang [1]. In order to highlight the difference and similarity between PEMM and PEMvM, we use notation that follows closely that of Chen, Prentice and Wang [1]. We denote the complete protein abundance matrix as $\mathbf{X} = (x_{ij})$, where x_{ij} represents the jth feature (i.e. protein) for the ith experiment where $i = 1, ..., n$ and $j = 1, ..., p$. It is assumed that $\mathbf{X}$ has a multivariate normal distribution. The missing indicator matrix is denoted as $\mathbf{M} = (m_{ij})$, where $m_{ij} = 0$ if x_{ij} is observed, and $m_{ij} = 1$ if x_{ij} is missing. $\mathbf{X}_{i,obs} = \{x_{ij} : j \in \mathbf{O}_i\}$ where $\mathbf{O}_i$ represents the index set of proteins being observed in experiment i. Then, the missing data mechanism in PEMM is:

$$\begin{aligned} P(\mathbf{M}|\mathbf{X}, \gamma_1, \gamma_2) &= \prod_{i,j} P(m_{ij} = 1|x_{ij}, \gamma_1, \gamma_2) \\ &= \prod_{i,j} \min(e^{\gamma_1 + \gamma_2 x_{ij}}, 1). \end{aligned} \tag{1}$$

If γ_2 is non-zero, its missing mechanism is NMAR. The negative γ_2 implies that the probability of not observing protein abundance increases as the protein abundance decreases. It is also assumed that a missing data mechanism of a protein abundance does not depend on the abundances of other proteins but only its own abundance. In (1), we decide not to include covariates (i.e. protein size) which may influence the probability of missing data, but does not depend on x_{ij} since such covariates do not contribute to the calculation of the maximum penalized likelihood estimates.

Then, with known missing data mechanism $\mathbf{\Gamma} = (\gamma_1, \gamma_2)$, the maximum likelihood estimation of $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}})$ is the following:

$$\begin{aligned} (\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}}) &= \text{arg max}_{\boldsymbol{\mu}, \mathbf{\Sigma}}\ L_{\mathbf{\Gamma}}(\boldsymbol{\mu}, \mathbf{\Sigma}) \\ &= \text{arg max}_{\boldsymbol{\mu}, \mathbf{\Sigma}} \{\sum_{i=1}^{n} \log f(\mathbf{X}_{i,obs}, \mathbf{M}_i; \boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{\Gamma}) - P(\mathbf{\Sigma})\}. \end{aligned} \tag{2}$$

where $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ are the mean and covariance matrixes of protein abundances, and $P(\mathbf{\Sigma})$ is a penalty term. $P(\mathbf{\Sigma}) = \lambda \sum_l 1/d_l + K \log(\prod_l d_l)$, where $\lambda > 0$ and $K > 0$ are penalty parameters and $\{d_l\}_{l=1}^{p}$ are the eigenvalues of $\mathbf{\Sigma}$. The choice of penalty parameters are $\lambda = 5$ and $K = 5$ based on previous numerical studies [1].

Given λ, K, and $\mathbf{\Gamma}$, we estimate $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}})$ using the following PEMM algorithm:

- Initialization. Set $\boldsymbol{\mu}^{(0)} = \bar{\mathbf{X}}$ and $\mathbf{\Sigma}^{(0)} = (n + K)^{-1}(n\mathbf{S}_{\mathbf{x}} + \lambda^{(0)}\mathbf{I})$ where $\bar{\mathbf{X}}$ and $\mathbf{S}_{\mathbf{x}}$ are the sample mean and the sample covariance based on available cases. $\lambda^{(0)}$ is the smallest positive that satisfies $\lambda^{(0)} \geq \lambda$ and the minimum eigenvalue of the matrix $n\mathbf{S}_{\mathbf{x}} + \lambda^{(0)}\mathbf{I}$ is positive.
- Iterate until the relative difference between $(b-1)^{th}$ and $(b)^{th}$ iterations is less than *tol*:
  - E-step: Given $(\boldsymbol{\mu}^{(b-1)}, \mathbf{\Sigma}^{(b-1)})$, estimate the conditional expectation of the sufficient statistics:
    $\mathbf{A}^{(b)}_{i:mis,mis} = \mathbf{\Sigma}^{(b-1)}_{i:mis,mis}$
    $-\mathbf{\Sigma}^{(b-1)}_{i:mis,obs}(\mathbf{\Sigma}^{(b-1)}_{i:obs,obs})^{-1}\mathbf{\Sigma}^{(b-1)}_{i:obs,mis}$,
    $\hat{\mathbf{X}}^{(b)}_{i,obs} = \hat{\mathbf{X}}_{i,obs}$,
    $\hat{\mathbf{X}}^{(b)}_{i,mis} = \boldsymbol{\mu}^{(b-1)}_{i,mis} + \mathbf{\Sigma}^{(b-1)}_{i:mis,obs}(\mathbf{\Sigma}^{(b-1)}_{i:obs,obs})^{-1}$
    $\times(\hat{\mathbf{X}}^{(b-1)}_{i,obs} - \boldsymbol{\mu}^{(b-1)}_{i,obs}) + \gamma_2 \mathbf{A}^{(b)}_{i:mis,mis} \cdot \mathbf{1}$.
  - M-step: Obtain the maximum penalized likelihood estimates:
    $\boldsymbol{\mu}^{(b)} = n^{-1} \sum_i \hat{\mathbf{X}}^{(b)}_i$,
    $\mathbf{\Sigma}^{(b)} = (n + K)^{(-1)}$
    $\times(\sum_i((\hat{\mathbf{X}}^{(b)}_i - \boldsymbol{\mu}^{(b)})(\hat{\mathbf{X}}^{(b)}_i - \boldsymbol{\mu}^{(b)})^T + \mathbf{A}^{(b)}_i) + \lambda^{(b)}\mathbf{I})$,
    where $\lambda^{(b)} < \lambda$ is chosen to be the smallest value which makes $\mathbf{\Sigma}^{(b)}$ positive-definite.
- Let $\boldsymbol{\mu} = \boldsymbol{\mu}^{(b)}$ and $\mathbf{\Sigma} = \mathbf{\Sigma}^{(b)}$.

2.2 PEMvM

To properly handle multiple missing data mechanisms across laboratories, we propose PEMvM (Penalized EM algorithm incorporating varying non-ignorable missing mechanism). Assuming experiments from the same laboratory were performed under similar condition, we group experiments by their laboratories. We let g_i be a group index of i^{th} experiment where there are q groups of experiments and $q \leq n$. To address different missing data mechanisms across groups, we define the missing data mechanisms in PEMvM as the following:

$$\begin{aligned} P(\mathbf{M}|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \prod_{i,j} P(m_{ij} = 1|x_{ij}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= \prod_{i,j} \min(e^{\alpha_{g_i} + \beta_{g_i} x_{ij}}, 1), \end{aligned} \tag{3}$$

where g_i is a group index of experiment i determined prior to analysis. We assume that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are known. In (3), each group g_i has unique missing data mechanism parameters. When $\gamma_1 = \alpha_1 = ... = \alpha_q$ and $\gamma_2 = \beta_1 = ... = \beta_q$, (1) and (3) are the same. Thus, for a common missing data mechanism case, PEMvM is reduced to PEMM. Similarly, if $\beta_{g_i} < 0$, the missing mechanism for experiment i is NMAR. Then, we obtain the maximum likelihood estimation of $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ given the missing data mechanisms, $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$:

(4)
$$\begin{aligned}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) &= \arg\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \; L_{\boldsymbol{\theta}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \arg\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \{\sum_{i=1}^{n} \log f(\mathbf{X}_{i,obs}, \mathbf{M}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\theta}) - P(\boldsymbol{\Sigma})\}.\end{aligned}$$

We use the same penalty term as the PEMM algorithm to ensure non-singular covariance matrix and concentration matrix ($\lambda = 5, K = 5$). The same choice of penalty term was used for both simulated and real data. Given λ, K, and $\boldsymbol{\theta}$, we estimate ($\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$) using the following PEMvM algorithm:

- Initialization. Set $\boldsymbol{\mu}^{(0)} = \bar{\mathbf{X}}$ and $\boldsymbol{\Sigma}^{(0)} = (n + K)^{-1}(n\mathbf{S}_{\mathbf{x}} + \lambda^{(0)}\mathbf{I})$ where $\bar{\mathbf{X}}$ and $\mathbf{S}_{\mathbf{x}}$ are the sample mean and the sample covariance based on available cases. $\lambda^{(0)}$ is the smallest positive that satisfies $\lambda^{(0)} \geq \lambda$ and the minimum eigenvalue of the matrix $n\mathbf{S}_{\mathbf{x}} + \lambda^{(0)}\mathbf{I}$ is positive.
- Iterate until the relative difference between $(b-1)^{th}$ and $(b)^{th}$ iterations is less than *tol*:
  - E-step: Given $(\boldsymbol{\mu}^{(b-1)}, \boldsymbol{\Sigma}^{(b-1)})$, estimate the conditional expectation of the sufficient statistics:
    $\mathbf{A}^{(b)}_{i:mis,mis} = \boldsymbol{\Sigma}^{(b-1)}_{i:mis,mis}$
    $-\boldsymbol{\Sigma}^{(b-1)}_{i:mis,obs}(\boldsymbol{\Sigma}^{(b-1)}_{i:obs,obs})^{-1}\boldsymbol{\Sigma}^{(b-1)}_{i:obs,mis}$,
    $\hat{\mathbf{X}}^{(b)}_{i,obs} = \hat{\mathbf{X}}_{i,obs}$,
    $\hat{\mathbf{X}}^{(b)}_{i,mis} = \boldsymbol{\mu}^{(b-1)}_{i,mis} + \boldsymbol{\Sigma}^{(b-1)}_{i:mis,obs}(\boldsymbol{\Sigma}^{(b-1)}_{i:obs,obs})^{-1}$
    $\times(\hat{\mathbf{X}}^{(b-1)}_{i,obs} - \boldsymbol{\mu}^{(b-1)}_{i,obs}) + \beta_{g_i}\mathbf{A}^{(b)}_{i:mis,mis} \cdot \mathbf{1}$.
  - M-step: Obtain the maximum penalized likelihood estimates:
    $\boldsymbol{\mu}^{(b)} = n^{-1}\sum_i \hat{\mathbf{X}}^{(b)}_i$,
    $\boldsymbol{\Sigma}^{(b)} = (n + K)^{(-1)}$
    $\times(\sum_i((\hat{\mathbf{X}}^{(b)}_i - \boldsymbol{\mu}^{(b)})(\hat{\mathbf{X}}^{(b)}_i - \boldsymbol{\mu}^{(b)})^T + \mathbf{A}^{(b)}_i) + \lambda^{(b)}\mathbf{I})$,
    where $\lambda^{(b)} < \lambda$ is chosen to be the smallest value which makes $\boldsymbol{\Sigma}^{(b)}$ positive-definite.
- Let $\boldsymbol{\mu} = \boldsymbol{\mu}^{(b)}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(b)}$.

In the E-step, PEMvM uses different missing data mechanisms, β_{g_i}, for experiments from different groups, while PEMM uses a common γ_2 for all experiments. This change in the missing data mechanism only affects the E-step, but does not interfere with the M-step. Thus, the M-step is the same for both PEMM and PEMvM. The proofs of PEMvM are unnecessary given their similarity to PEMM.

We also impute the missing data using missing value estimation at the final iteration b in the E-step of the PEMvM algorithm:

(5)
$$\begin{aligned}\hat{\mathbf{X}}^{(b)}_{i,mis} &= \boldsymbol{\mu}^{(b-1)}_{i,mis} + \boldsymbol{\Sigma}^{(b-1)}_{i:mis,obs}(\boldsymbol{\Sigma}^{(b-1)}_{i:obs,obs})^{-1} \\ &\quad \times(\hat{\mathbf{X}}^{(b-1)}_{i,obs} - \boldsymbol{\mu}^{(b-1)}_{i,obs}) + \beta_{g_i}\mathbf{A}^{(b)}_{i:mis,mis} \cdot \mathbf{1}.\end{aligned}$$

=where $\hat{\mathbf{X}}^{(b)}_{i,mis}$ is the conditional expectation of missing data given the non-ignorable missing data mechanisms and the observed data. The R codes are available at https://github.com/soyoungryu/PEMvM_usingR.

3. RESULTS AND DISCUSSION

3.1 Simulation results

We simulated multivariate normal data with varying missing data mechanisms. The missing probability was calculated using Equation (3). We let $\beta_{g_i} = 0.2$ where $i = 1$,..., $\lceil n/2 \rceil$ and $\beta_{g_i} = w$ where $i = \lceil n/2 \rceil + 1, ..., n$. We varied w to be either 0.4 or 0.6 in order to investigate the association between PEMvM performance and missing data mechanism differences between groups. We set α_{g_i} such that the missing rates approximately ranged between 35% and 45%. μ_j was randomly sampled from the uniform distribution with a minimum value of 3 and a maximum value of 8. We let $\Sigma_{jj} = 1$ and $\Sigma_{jj'}$ with $j \neq j'$ sampled from either zero or $N(0.5, 0.1^2)$. Given $p = 30$, we considered the data with $n < p$ and $n \geq p$ by varying n to be $10, 20, 30, 40$, or 50. We compared the performance between PEMvM and PEMM using the mean standard errors between estimated values and true values. For each simulation set, we performed 1,000 simulations and reported the average values of simulation results with the corresponding standard errors. In PEMvM, the given $(\beta_{g_1}, ..., \beta_{g_n})$ were used assuming that missing mechanisms were known. Similarly, γ_2 in PEMM was set to be a mean of $(\beta_{g_1}, ..., \beta_{g_n})$.

When multiple non-ignorable missing data mechanisms were present, PEMvM performed better than PEMM in terms of accurate missing value imputation and mean estimation (Table 1 and 2). But these two algorithms performed similarly in covariance estimations with a small improvement for PEMvM (Table 3). More specifically, the MSEs of imputed values decreased for both PEMM and PEMvM as the sample size increased; however, the MSEs of imputed values were smaller using a PEMvM approach (Table 1). The superior performance of PEMvM in missing data imputation was more noticeable when missing data mechanism differences between groups were larger. When $\beta_1 = 0.2$ and $\beta_2 = 0.4$, we were able to decrease the MSEs of imputed values by 6.69% to 17.36% using PEMvM. When $\beta_1 = 0.2$ and $\beta_2 = 0.6$, the MSEs of imputed values were decreased by 14.57% to 26.64% using PEMvM. For mean estimations, similar patterns were observed (Table 2). When the difference between β_1 and β_2 was 0.2, the MSEs of mean estimations were improved by 10.50% to 18.78% using PEMvM. When the difference between β_1 and β_2 was 0.4, the MSEs of mean estimations were improved by 9.09% to 18.26% using PEMvM. For both cases, the MSEs of covariance estimations between the two algorithms differ by less than 1% with PEMvM performing slightly better. In summary, as the missing data mechanism difference became larger between groups, the benefit of PEMvM was more pronounced,

Table 1. Average mean squared errors (MSEs) between imputed missing values and true values over 1,000 simulations. Standard errors were reported in parentheses. A relative difference represents a MSE difference of imputed data using two algorithms divided by the MSE of PEMM. The negative relative difference means a decrease in MSE using a PEMvM approach

$\beta_1 = 0.2$ vs. $\beta_2 = 0.4$	n=10	n=20	n=30	n=40	n=50
PEMM	1.7604 (0.0177)	1.5128 (0.0130)	1.2072 (0.0066)	1.0685 (0.0038)	0.9971 (0.0027)
PEMvM	1.4683 (0.0117)	1.2501 (0.0076)	1.0832 (0.0041)	0.9971 (0.0027)	0.9661 (0.0023)
relative difference	-16.59%	-17.36%	-10.27%	-6.69%	-8.32%

$\beta_1 = 0.2$ vs. $\beta_2 = 0.6$	n=10	n=20	n=30	n=40	n=50
PEMM	1.6981 (0.0557)	1.4496 (0.0089)	1.3169 (0.0065)	1.1805 (0.0041)	1.1699 (0.0040)
PEMvM	1.2458 (0.0256)	1.0887 (0.0043)	1.0198 (0.0034)	1.0084 (0.0027)	0.9658 (0.0023)
relative difference	-26.64%	-24.90%	-22.56%	-14.57%	-17.45%

Table 2. Average mean squared errors (MSEs) between $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{\mu}$ over 1,000 simulations. Standard errors were reported in parentheses. A relative difference represents a MSE difference of mean estimations using two algorithms divided by the MSE of PEMM. The negative relative difference means a decrease in MSE using a PEMvM approach

$\beta_1 = 0.2$ vs. $\beta_2 = 0.4$	n=10	n=20	n=30	n=40	n=50
PEMM	0.2291 (0.0034)	0.1419 (0.0023)	0.0811 (0.0014)	0.0500 (0.0007)	0.0435 (0.0007)
PEMvM	0.2018 (0.0028)	0.1152 (0.0017)	0.0679 (0.0010)	0.0447 (0.0005)	0.0370 (0.0004)
relative difference	-11.89%	-18.78%	-16.33%	-10.50%	-14.97%

$\beta_1 = 0.2$ vs. $\beta_2 = 0.6$	n=10	n=20	n=30	n=40	n=50
PEMM	0.1744 (0.0070)	0.0983 (0.0013)	0.0645 (0.0008)	0.0479 (0.0006)	0.0409 (0.0005)
PEMvM	0.1586 (0.0057)	0.0811 (0.0009)	0.0527 (0.0006)	0.0400 (0.0004)	0.0335 (0.0004)
relative difference	-9.09%	-17.47%	-18.26%	-16.56%	-18.17%

Table 3. Average mean squared errors (MSEs) between $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}$ over 1,000 simulations. Standard errors were reported in parentheses. A relative difference represents a MSE difference of mean estimations using two algorithms divided by the MSE of PEMM. The negative relative difference means a decrease in MSE using a PEMvM approach

$\beta_1 = 0.2$ vs. $\beta_2 = 0.4$	n=10	n=20	n=30	n=40	n=50
PEMM	0.0701 (0.0007)	0.0617 (0.0007)	0.0435 (0.0066)	0.0333 (0.0003)	0.0315 (0.0004)
PEMvM	0.0649 (0.0004)	0.0540 (0.0004)	0.0395 (0.0041)	0.0315 (0.0002)	0.0279 (0.0002)
relative difference	-0.30%	-0.51%	-0.33%	-0.17%	-0.33%

$\beta_1 = 0.2$ vs. $\beta_2 = 0.6$	n=10	n=20	n=30	n=40	n=50
PEMM	0.0635 (0.0014)	0.0498 (0.0004)	0.0418 (0.0065)	0.0315 (0.0002)	0.0297 (0.0002)
PEMvM	0.0604 (0.0011)	0.0452(0.0003)	0.0373 (0.0034)	0.0292 (0.0002)	0.0246 (0.0002)
relative difference	-0.18%	-0.31%	-0.34%	-0.20%	-0.20%

especially for missing value imputations and mean estimations. At various sample sizes, PEMvM had smaller MSEs of imputed value, $\hat{\boldsymbol{\mu}}$, and $\hat{\boldsymbol{\Sigma}}$ compared to PEMM.

3.2 Spiked-in human proteins in yeast

We applied PEMvM to 45 UPS1 (Universal Proteomics Standard Set 1) proteins that were previously used to investigate the performance of PEMM compared to other methods (i.e. available case analysis, k-nearest neighbor imputation, penalized EM algorithm with MAR assumption) [1]. In brief, UPS1 proteins were spiked in the yeast lysate samples at three different concentrations. Then, each protein mixture was analyzed by mass spectrometry three times at four different laboratories. The software *Sahale* [7] was used to measure protein abundance in these samples. The dataset was available in Chen, Prentice and Wang [1]. The median values of yeast protein abundances were used to normalize data since their concentrations were equal in all mixtures. There were protein mixtures with UPS1 protein concentrations of 20 fmol/μL, 6.7 fmol/μL and 2.2 fmol/μL. The missing rates for these sets were 9.8%, 23.7%, and 51.1% with the least missing rate for the highest UPS1 protein concentration. We labeled them as Mixture 1, 2, and 3, respectively. Since it was more adequate to use a dataset with the least missing rate (Mixture 1) as a reference dataset, we compared UPS1 proteins between Mixture 1 vs. Mixture 2 (Comparison I) and Mixture 1 vs. Mixture 3 (Comparison II). Since missing data mechanisms were unknown for this experiment, we estimated $\boldsymbol{\theta}_{\boldsymbol{g_i}} = (\alpha_{g_i}, \beta_{g_i})$ using available-case mean estimates of protein abundances and their missing percentages as shown in Chen et al. [2].

Since the mixtures were analyzed at multiple laboratories, we suspected that the protein abundance data generated at different laboratories had different missing data mechanisms. As shown in Table 4, the missing rates varied across laboratories. Moreover, different laboratories had different magnitudes of missing rate changes between mixtures. For example, the missing rate difference between Mixture 1 and 2 in Laboratory 3 was 18.52% (= 20.74%−2.22%), while their difference in Laboratory 2 was 6.67% (= 15.56% − 8.89%). The values of $\hat{\beta}_{g_i}$ ranged from 0.05 to 0.17.

When PEMvM was applied to this multi-laboratory data, PEMvM estimated (mean) protein abundances more accurately than PEMM (Table 5). The true UPS1 protein concentration differences between mixtures were used to indirectly measure the accuracy of means estimated by PEMM and PEMvM. We note that, due to varying ion efficiencies of peptides (i.e. fragments of proteins), it was not feasible to directly measure the actual protein abundances [3, 11]. However, since the same peptide sequences had the same ion efficiencies, it was possible to measure relative abundance differences between mixtures and use this information to measure the accuracy of means estimated by PEMM and PEMvM. When PEMvM was employed instead of PEMM, the MSEs of mean protein abundance differences between two mixtures decreased by 7.42% for Comparison I (Mixture 1 vs. 2) and by 1.28% for Comparison II (Mixture 1 vs. 3). The magnitudes of MSEs were quite different between Comparison I and II. This was expected since Mixture 3 had a missing rate of over 50%, thus making it harder to accurately estimate protein abundance differences. We also estimated the variance of protein abundance differences among UPS1 proteins. Since the true abundance differences between mixtures were the same for all 45 UPS1 proteins, it would be ideal if variance was similar to the MSE. As shown in Table 5, the MSEs were reasonably close to their variances.

PEMvM performed better than PEMM in terms of accurately measuring relative protein ratios in a presence of varying missing mechanisms; however, we want to note limitations of our proposed method. First, in our study, we assume that missing mechanisms are known; however, jointly estimating missing mechanisms and protein abundances is important, but is not an easy task due to computational limitation. Thus, future study on this topic is necessary to further improve the performances of both PEMM and PEMvM. Secondly, PEMvM must be applied with caution to the dataset that may have a common missing mechanism. In such cases, PEMvM will be reduced to PEMM if missing mechanisms can be accurately estimated; however, when missing data mechanisms cannot be estimated accurately, PEMvM may perform worse than PEMM owing to uncertainty in measuring missing data mechanisms.

Table 4. The missing rates of UPS1 proteins

	Mixture 1	Mixture 2	Mixture 3
Laboratory 1	9.63%	22.22%	42.96%
Laboratory 2	8.89%	15.56%	39.26%
Laboratory 3	2.22%	20.74%	54.07%
Laboratory 4	18.52%	36.30%	68.15%

Table 5. MSE errors of log_2 transformed mean protein ratios for Comparison I (Mixture 1 vs. 2) and Comparison II (Mixture 1 vs. 3)

Comparison I	MSE	Variance
PEMM	0.5418	0.5418
PEMvM	0.5016	0.5000
relative difference	-7.42%	-7.72%
Comparison II	MSE	Variance
PEMM	1.8406	1.6606
PEMvM	1.8170	1.6382
relative difference	-1.28%	-1.35%

4. CONCLUSION

The multi-laboratory consortium or large-scale proteomic studies can provide valuable information about robust protein biomarkers for diseases; however, careful attention is needed to properly analyze such datasets. Among many challenges in analyzing such datasets, we are interested in properly handling non-ignorable missing data mechanisms. In this paper, we extended the PEMM algorithm to incorporate varying missing data mechanisms across laboratories. Our algorithm PEMvM can also be used in large-scale proteomic studies, especially when the biological samples are analyzed over several years. For example, missing data patterns may vary after a major instrument tuning, even when the samples are analyzed by the same instrument. Thus, it would be necessary to consider the varying missing data mechanisms.

ACKNOWLEDGEMENTS

This publication was made possibly by a grant from the National Institute of General Medical Sciences (GM103440) from the National Institutes of Health.

Received 28 December 2017

REFERENCES

[1] Chen, L. S., Prentice, R. L. and Wang, P. (2014). A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation. *Biometrics* **70** 312–322. MR3258036

[2] Chen, L. S., Wang, J., Wang, X. and Wang, P. (2016). A Mixed-effects Model for Incomplete Data With Batch-Level Abundance-Dependent Missing-Data Mechanism. *ArXiv e-prints.*

[3] Fusaro, V. A., Mani, D. R., Mesirov, J. P. and Carr, S. A. (2009). Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat Biotechnol* **27** 190–198.

[4] Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*, 2nd ed. *Wiley series in probability and statistics*. Wiley, Hoboken, N.J. MR1925014

[5] Liu, H., Sadygov, R. G. and Yates, r. J. R. (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **76** 4193–4201.

[6] Luo, R., CM, C., WC, S. and H, Z. (2009). Bayesian Analysis of iTRAQ Data with Nonrandom Missingness: Identification of Differentially Expressed Proteins. *Statistics in Biosciences* **1** 228–245.

[7] Milac, T. I., Randolph, T. W. and Wang, P. (2012). Analyzing LC-MS/MS data by spectral count and ion abundance: two case studies. *Stat Interface* **5** 75–87. MR2896982

[8] Paulovich, A. G., Billheimer, D., Ham, A. J., Vega-Montoto, L., Rudnick, P. A., Tabb, D. L., Wang, P., Blackman, R. K., Bunk, D. M., Cardasis, H. L., Clauser, K. R., Kinsinger, C. R., Schilling, B., Tegeler, T. J., Variyath, A. M., Wang, M., Whiteaker, J. R., Zimmerman, L. J., Fenyo, D., Carr, S. A., Fisher, S. J., Gibson, B. W., Mesri, M., Neubert, T. A., Regnier, F. E., Rodriguez, H., Spiegelman, C., Stein, S. E., Tempst, P. and Liebler, D. C. (2010). Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol Cell Proteomics* **9** 242–254.

[9] Rubin, D. B. (1976). Inference and Missing Data. *Biometrika* **63** 581–592. MR0455196

[10] Ryu, S. Y., Qian, W. J., Camp, D. G., Smith, R. D., Tompkins, R. G., Davis, R. W. and Xiao, W. (2014). Detecting differential protein expression in large-scale population proteomics. *Bioinformatics* **30** 2741–2746.

[11] Tang, H., Arnold, R. J., Alves, P., Xun, Z., Clemmer, D. E., Novotny, M. V., Reilly, J. P. and Radivojac, P. (2006). A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics* **22** e481–e488.

So Young Ryu
School of Community Health Sciences
University of Nevada Reno
1664 N. Virginia Street
Reno, NV 89557
USA
E-mail address: soyoungr@unr.edu