Biostatistics (2018) **00**, 00, *pp.* 1–18
doi:10.1093/biostatistics/kxy022

Using multivariate mixed-effects selection models for analyzing batch-processed proteomics data with non-ignorable missingness

JIEBIAO WANG[†]

Department of Public Health Sciences, University of Chicago, 5841 S. Maryland Ave., Chicago, IL 60637, USA

PEI WANG

Department of Genetics and Genomics Sciences, Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, 770 Lexington Avenue, New York, NY 10065, USA

DONALD HEDEKER, LIN S. CHEN[*]

Department of Public Health Sciences, University of Chicago, 5841 S. Maryland Ave., Chicago, IL 60637, USA

lchen@health.bsd.uchicago.edu

SUMMARY

In quantitative proteomics, mass tag labeling techniques have been widely adopted in mass spectrometry experiments. These techniques allow peptides (short amino acid sequences) and proteins from multiple samples of a batch being detected and quantified in a single experiment, and as such greatly improve the efficiency of protein profiling. However, the batch-processing of samples also results in severe batch effects and non-ignorable missing data occurring at the batch level. Motivated by the breast cancer proteomic data from the Clinical Proteomic Tumor Analysis Consortium, in this work, we developed two tailored multivariate MIxed-effects SElection models (mvMISE) to jointly analyze multiple correlated peptides/proteins in labeled proteomics data, considering the batch effects and the non-ignorable missingness. By taking a multivariate approach, we can borrow information across multiple peptides of the same protein or multiple proteins from the same biological pathway, and thus achieve better statistical efficiency and biological interpretation. These two different models account for different correlation structures among a group of peptides or proteins. Specifically, to model multiple peptides from the same protein, we employed a factor-analytic random effects structure to characterize the high and similar correlations among peptides. To model biological dependence among multiple proteins in a functional pathway, we introduced a graphical lasso penalty on the error precision matrix, and implemented an efficient algorithm based on the alternating direction method of multipliers. Simulations demonstrated the advantages of the proposed models. Applying the proposed methods to the motivating data set, we identified phosphoproteins and biological pathways that showed different activity patterns in triple negative breast tumors versus other breast tumors. The proposed methods can also be applied to other high-dimensional multivariate analyses based on clustered data with or without non-ignorable missingness.

[*]To whom correspondence should be addressed.
[†]Present address: Department of Statistics and Data Science, Carnegie Mellon University, 4909 Frew St, Pittsburgh, PA 15213, USA

Keywords: Alternating direction method of multipliers; Expectation-maximization algorithm; Graphical lasso; Missing not at random; Multivariate mixed-effects models; Selection model; Proteomics.

1. INTRODUCTION

In quantitative proteomics research, mass spectrometry (MS) experiments are widely used for shotgun profiling of proteins and peptides. In these experiments, proteins are digested into smaller sequences of amino acids, i.e., peptides. After the abundances of peptides are derived from MS experiments, one may analyze each peptide as the analysis unit or summarize multiple peptides in a protein and analyze each protein as the analysis unit. Traditional shotgun MS experiments process each sample one by one, and thus greatly limit the scale of quantitative proteomics research. More recently, mass tag labeling techniques such as isobaric tags for relative and absolute quantitation (iTRAQ) and tandem mass tags (TMT), have been widely adopted in MS experiments (Wiese *and others*, 2007). These techniques allow proteins/peptides from multiple samples of a batch being quantified in a single MS experiment, and thus greatly enhance the efficiency of data generation. However, the batch processing of samples results in severe batch effects. Additionally, in the general MS experiments, the lower the abundance of a peptide, the more likely the peptide is missing. The batch-processing of data also results in the missing data occurring at the batch level, i.e., a peptide abundance is either all observed or all missing in the samples processed by the same experiment. The lower the average abundance of a peptide in a batch of samples is, the more likely the peptide would be missing in all the samples of the batch.

To correct for batch effects, a shared reference sample is often included in all batches. Conventional analyses are performed based on the observed relative abundances of peptides in the target samples to the common reference. However, since the target and reference samples are often subject to very different experimental variations (Karp *and others*, 2010), analyses of relative abundances of peptides/proteins cannot fully correct for batch effects. Moreover, most existing analyses are based on only the observed abundances and ignore the missing data. Here, the missing data are non-ignorable (Little and Rubin, 2002) because the missing probability of a peptide/protein in a batch largely depends on its average abundances of all samples in the batch. Ignoring the missing data may lead to biased estimation and inference.

To analyze labeled (batch-processed) proteomics data, Chen *and others* (2017b) recently proposed a univariate mixed-effects model accounting for the batch-processing design and the batch-level non-ignorable missing-data mechanism. This univariate model can be used to directly analyze each individual peptide. When using this model to analyze each protein, one may take the average abundance level of peptides from the same protein as the protein abundance level. However, this strategy may not be optimal. Those summary abundances of proteins could be dominated by peptides of high abundances, and the information in low abundance peptides is largely ignored. Indeed, for unlabeled proteomics experiments (in which samples are processed one by one), Clough *and others* (2009) have suggested that by jointly modeling individual peptides of the same protein, one can gain improved precision and power relative to analyses based on averaging peptide abundances. For labeled proteomics experiments, to our knowledge, there is no multivariate analysis tool available to model peptide level information directly while taking into consideration all aforementioned properties of labeled proteomics experiments. In this work, we aim to bridge this gap and develop multivariate approaches to better aggregate information from peptides of the same proteins and to jointly analyze multiple proteins from a pathway. The proposed multivariate methods and tools complement the univariate approaches and aim to detect individually weak-to-moderate and collective strong effects.

Herein, we consider the problem of jointly analyzing multiple correlated outcomes or response variables in labeled proteomics data. Note that in the current work, "outcome" refers to response variable, not clinical outcome. A desirable model should consider the batch design, the non-ignorable missing-data mechanism, and the correlations among multiple response variables. Existing methods have been proposed

for multivariate analysis within the mixed-effects model framework. Roy and Lin (2002) proposed a latent variable selection model for analyzing multivariate longitudinal data with missing values caused by dropouts. The term "selection model" refers to a class of statistical models in missing data analysis. Generally, the likelihood function is based on a model for the joint distribution of the data and the missing data mechanism. In selection models, this joint distribution is factored into a distribution for the data if there were no missing values, and a model of the missing data indicators given the data (Little and Rubin, 2002). Liu and Hedeker (2006) developed a selection model for bivariate responses with missing values occurring sporadically. However, these models cannot be directly applied to labeled proteomics data with the unique clustered non-ignorable missingness. Moreover, most existing methods can only analyze multivariate response variables of low dimensionality, whereas in proteomics data, the number of peptides in a protein and the number of proteins in a pathway can be up to hundreds and may exceed the sample size. Additionally, standard procedures for mixed-effects models after imputations of missing values are not sufficient in this setting either, because most existing imputation methods in the standard software packages are developed to analyze ignorable missing data and do not account for the unique non-ignorable missing data mechanism in labeled proteomics data.

These challenges motivated us to develop multivariate models incorporating the batch design and the batch-level (or cluster-level) non-ignorable missingness. We termed them as the multivariate MIxed-effects SElection (mvMISE) models. Our method employed tailored modeling for different correlation structures of different types of outcomes and can achieve efficient parameter estimation for high-dimensional outcomes. Specifically, we developed two multivariate models: Model I—mvMISE$_b$—for jointly analyzing multiple peptides from each protein, and Model II—mvMISE$_e$—for jointly analyzing multiple proteins from *a priori* defined pathways. In both models, we considered estimation feasibility and computational efficiency. We applied our methods to an iTRAQ-based breast cancer proteomic data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) (Mertins *and others*, 2016). The proposed models can be extended to analyze general clustered multivariate data with or without (non-ignorable) missing data.

2. Methods

2.1. *A multivariate mixed-effects model for batch-processed (or clustered) data*

Let $\mathbf{Y}_i$ be an $n_i \times K$ matrix, denoting the abundance levels of K peptides in a protein (or K proteins in a pathway) from the i-th experiment (i.e., the i-th batch of samples), $i = 1, 2, \ldots, N$. In our multivariate analysis, K is the number of correlated response variables. Let n_i be the number of samples in the i-th batch including the reference sample. In the motivating CPTAC data, every three randomly selected target tumor samples and one common reference sample were grouped into a batch and were analyzed by a four-plex (i.e., four-channel) iTRAQ experiment. A total of 108 tumor samples were analyzed in 36 iTRAQ experiments. The samples in the same batch were processed together, and the quantitations of abundance levels from these samples were related. Let the vector $\mathbf{y}_i = \text{vec}(\mathbf{Y}_i)$ denote all the abundance levels of K-variate outcomes in the i-th batch. A natural model for the batch-processed data is a mixed-effects model with multivariate abundance levels as the outcome variables:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \tag{2.1}$$

where $\mathbf{X}_i$ is an $n_iK \times p$ design matrix with fixed effects $\boldsymbol{\alpha}$, and $\mathbf{Z}_i$ is an $n_iK \times q$ design matrix with random effects $\mathbf{b}_i$. We assume that the random effects $\mathbf{b}_i$ follow a normal distribution $\mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{D})$, where the matrix $\mathbf{D}$ captures the covariances of random effects. Let $\mathbf{e}_i = \text{vec}(\mathbf{E}_i)$, where $\mathbf{E}_i$ is the $n_i \times K$ error matrix. We assume that $\mathbf{E}_i$ follows a matrix normal distribution $\mathbf{E}_i \sim MN_{n_i,K}(\mathbf{0}, \mathbf{S}_i, \boldsymbol{\Sigma})$, and thus $\mathbf{e}_i$ follows a multivariate normal distribution $\mathbf{e}_i \sim N_{n_iK}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{S}_i)$, where $\otimes$ denotes the Kronecker product. The matrix $\mathbf{S}_i = \text{diag}\left(\sigma_0^2, \sigma^2, \ldots, \sigma^2\right)$ is an $n_i \times n_i$ diagonal matrix with the first diagonal element (σ_0^2)

corresponding to the variance of the common reference sample, and the rest of the diagonal elements (σ^2) being the variance of the target tumor samples. The variance for the reference is different and generally smaller than that of other tumor samples because the reference sample is created by combining multiple tumor samples. The matrix $\mathbf{\Sigma}$ captures the error (or unexplained) covariances among K response variables. The model can be used to estimate the common fixed effects averaged across K outcome variables and the outcome-specific effects. To estimate the common fixed effects, one may let $\mathbf{X}_i = \mathbf{1}_K \otimes \mathbf{X}_i^*$, where $\mathbf{1}_K$ is a vector of 1s and $\mathbf{X}_i^*$ is an $n_i \times p$ covariates matrix shared among K outcomes (for example, the tumor subtype indicators). To estimate outcome-specific effects, one may include outcome-specific covariates or the interactions of the predictors of interest and the indicators of outcomes in the design matrix, for example, let $\mathbf{X}_i = \mathbf{I}_K \otimes \mathbf{X}_i^*$ with $\mathbf{I}_K$ being an identity matrix.

2.2. *The non-ignorable batch-level missing-data mechanism*

A major challenge in analyzing proteomic data is the substantial amount of non-ignorable missing data. The probability of abundance levels of a peptide/protein being missing largely depends on the values themselves, and as such the missing data are non-ignorable. For batch-processed proteomics data, another complication is that the abundances of a peptide are generally either all missing or all observed in the samples from the same batch (i.e., the same experiment), with missing probability depending on the peptide's average abundances in the batch. In other words, not all the K-variate response variables (K outcomes) are fully observed in all experiments. The number of observed peptides is batch-dependent (or experiment-dependent). Chen *and others* (2017b) termed it as the "batch-level abundance-dependent" missing-data mechanism and propose to model it with an exponential function of the abundances. The exponential function can be naturally integrated with the normal density function, facilitating the estimation and computation of the employed Expectation-Maximization (EM) algorithm. Following this idea, we modeled the missing probability for each of the k-th outcomes ($k = 1, 2, \ldots, K$) in the i-th batch as a function of the average abundances of outcome k in batch i and other covariates

$$\Pr\left(r_{ik} = 1 | \mathbf{y}_{ik}\right) = g^{-1}\left(\phi_0 + \phi_1 / n_i \cdot \mathbf{1}_{n_i}^{'} \mathbf{y}_{ik} + \phi_2^{'} \mathbf{c}_{ik}\right), \tag{2.2}$$

where $\mathbf{y}_{ik}$ is a vector of abundances of the k-th outcome for the samples from the i-th batch, $g(\cdot)$ is the link function, r_{ik} is a missing indicator with $r_{ik} = 1$ if all of the n_i values in $\mathbf{y}_{ik}$ are missing, and $\mathbf{c}_{ik}$ denotes the average of covariates for the k-th outcome of the i-th batch, if there is any. The parameters ϕ_0, ϕ_1 and ϕ_2 control the missing-data mechanism. When $\phi_1 = 0$, the missing probability does not depend on the abundance values and the missing data are missing at random (Little and Rubin, 2002). The multivariate mixed-effects model (2.1) and the missing-data model (2.2) compose the proposed multivariate mvMISE models.

In model (2.2), using a logit or probit link function can be computationally prohibitive when analyzing high-dimensional data. Following Chen *and others* (2017b), we used a log link function

$$\Pr\left(r_{ik} = 1 | \mathbf{y}_{ik}\right) = \exp\left(\phi_0 + \phi_1 / n_i \cdot \mathbf{1}_{n_i}^{'} \mathbf{y}_{ik}\right). \tag{2.3}$$

Note that there are no outcome-specific covariates in the motivating data.

2.3. *The mvMISE models with different correlation structures tailored for different high-dimensional outcomes*

In order to model the correlations among multivariate or even high-dimensional outcomes, it is important to understand the nature of correlation structures. In this section, we considered two distinct applications

in proteomics data analyses, and developed tailored models for different correlation structures in different applications. These models consider the scalability and computation feasibility for jointly analyzing high-dimensional outcomes.

2.3.1. *The mvMISE model with correlated random effects (mvMISE$_b$) for analyzing multiple peptides from a protein.* In the MS proteomics experiments, large proteins are digested into smaller peptides, and peptides are measured by the MS instruments. Multiple peptides from the same protein have very similar abundance levels and are highly correlated. To model this type of correlations, we allow peptide-specific random effects to be correlated using a non-diagonal covariance matrix $\mathbf{D}$. Additionally, we assume $\mathbf{\Sigma} = \mathbf{I}_K$. That is, the error terms of multiple peptides from the same protein are uncorrelated after accounting for the correlations of random effects and the covariates, i.e., there are no unexplained correlations among peptides in a protein.

When the number of peptides (K) in a protein is large, it becomes computationally prohibitive to estimate the above models with an unstructured $\mathbf{D}$ matrix. Due to the highly correlated nature of those peptides from the same protein, we propose to employ a factor-analytic random-effects structure for the correlated random effects (Liu and Hedeker, 2006). We term this model as the mvMISE model with correlated random effects, mvMISE$_b$, and write it as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\boldsymbol{\tau}\, b_i + \mathbf{e}_i, \tag{2.4}$$

where $\mathbf{Z}_i = \mathbf{I}_K \otimes \mathbf{1}_{n_i}$, $\boldsymbol{\tau}$ is a $K \times 1$ vector for the peptide-specific variance components corresponding to the random effect b_i, and b_i is a standard normal random variable. In model (2.4), only K rather than $K(K+1)/2$ parameters need to be estimated for the covariance matrix of random effects, which substantially speeds up the computation. The model assumes a latent variable underlying the peptides in the same proteins and implies correlations equal one among random effects. Since the peptides from a protein are segments of the same molecule, the correlations are not only high but also similar in magnitude. This simplifying assumption is reasonable.

2.3.2. *The mvMISE model with correlated error terms (mvMISE$_e$) for pathway analysis.* In protein pathway analysis, one jointly analyzes multiple proteins in *a priori* defined functional pathways or protein sets. One may first obtain the protein abundance levels by averaging the peptide abundances mapped to each protein, and then treat multiple protein abundances in a pathway as the multivariate outcome measures. A major challenge in this analysis is to model the biological correlations among multiple proteins. Those correlations are often unstructured. Herein, we propose to model the unstructured biological correlations among proteins via the error covariance matrix Σ, while assuming the random effects are independent among outcomes. We term this model as the mvMISE model with correlated error terms, mvMISE$_e$.

Protein abundances in a functional pathway can be generally correlated, while their inter-dependence (i.e., network or conditional correlation) structures are often sparse (Baladandayuthapani *and others*, 2014). Therefore, we introduce a graphical lasso penalty on the error precision matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ (Danaher *and others*, 2014). We propose to obtain the estimates by

$$\text{maximize}_{\mathbf{\Theta}} \quad 2l\left(\mathbf{\Theta}\right) - \lambda N \left|\mathbf{\Theta}\right|_1, \tag{2.5}$$

where λ is a tuning parameter and $\left|\mathbf{\Theta}\right|_1 = \sum_{i \neq j} \left|\theta_{ij}\right|$. The graphical lasso penalty term ($\lambda \sum_{i \neq j} \left|\theta_{ij}\right|$) imposes regularization on the off-diagonal elements of the error precision matrix and as such ensures a sparse dependence structure. This will greatly facilitate the estimation of covariance and precision matrices for high-dimensional outcomes.

3. MODEL ESTIMATION

3.1. *An EM algorithm for the mvMISE$_b$ model*

In this section, we derive an EM algorithm for the proposed mvMISE$_b$ model with correlated outcome-specific random effects and an exponential missing-data mechanism function. The maximum likelihood estimation consists of iterating between two steps, the E-step and the M-step. The conditional expectations of the sufficient statistics in the E-step are derived and presented in Appendix A of the supplementary material available at *Biostatistics* online. In short, the non-ignorable missing-data mechanism function in (2.3) is integrated with the multivariate normal distribution function and imposes a bias-correction for the estimated conditional expectation of the missing values. In the M-step, we maximize the expected complete-data log-likelihood $Q\left(\boldsymbol{\gamma}\,|\,\boldsymbol{\gamma}^{(t)}\right)$ with respect to each parameter, $\boldsymbol{\gamma} = \{\boldsymbol{\alpha}, \boldsymbol{\tau}, \sigma_0^2, \sigma^2, \phi_0, \phi_1\}$, to obtain the maximum likelihood estimates (MLEs).

Let $\mathbf{y}_i$ denote the complete data $\mathbf{y}_i = \left(\mathbf{y}_i^{o'}, \mathbf{y}_i^{m'}\right)'$, where $\mathbf{y}_i^o$ and $\mathbf{y}_i^m$ are the observed and missing values, respectively. The matrices $\mathbf{X}_i^o$ and $\mathbf{X}_i^m$ are the design matrices corresponding to $\mathbf{y}_i^o$ and $\mathbf{y}_i^m$, and both matrices are fully observed. For the mvMISE$_b$ model in (2.4), let $\boldsymbol{\Sigma}_i = \mathbf{Z}_i \boldsymbol{\tau}\, \boldsymbol{\tau}' \mathbf{Z}_i' + \mathbf{R}_i$, where $\mathbf{R}_i = \mathbf{I}_K \otimes \mathbf{S}_i$. In each iteration of the M-step, we obtain the following estimates. The estimate of $\boldsymbol{\tau}$ for random effects is given by

$$\boldsymbol{\tau}^{(t+1)} = \left\{\sum_{i=1}^{N} \mathrm{E}^2\left(b_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) \mathbf{Z}_i' \mathbf{R}_i^{(t)^{-1}} \mathbf{Z}_i\right\}^{-1} \times \sum_{i=1}^{N} \left[\mathrm{E}\left(b_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) \mathbf{Z}_i' \mathbf{R}_i^{(t)^{-1}} \left\{\mathrm{E}\left(\mathbf{y}_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) - \mathbf{X}_i \boldsymbol{\alpha}^{(t)}\right\}\right].$$

The estimate of the fixed effects is

$$\boldsymbol{\alpha}^{(t+1)} = \left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{R}_i^{(t)^{-1}} \mathbf{X}_i\right)^{-1} \sum_{i=1}^{N} \left[\mathbf{X}_i' \mathbf{R}_i^{(t)^{-1}} \left\{\mathrm{E}\left(\mathbf{y}_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) - \mathbf{Z}_i \boldsymbol{\tau}^{(t)} \mathrm{E}\left(b_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right)\right\}\right].$$

The variance estimates of the error term are

$$\sigma_0^{2(t+1)} = \frac{1}{NK} \sum_{i=1}^{N} \mathbf{V}_{i,11}^{(t)} \text{ and } \quad \sigma^{2(t+1)} = \frac{1}{K\left(\sum_{i=1}^{N} n_i - N\right)} \sum_{i=1}^{N} \mathrm{tr}\left(\mathbf{V}_{i,-1,-1}^{(t)}\right),$$

where $\mathbf{V}_i^{(t)} = \mathrm{E}\left(\mathbf{e}_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) \mathrm{E}\left(\mathbf{e}_i'|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) + \mathrm{var}\left(\mathbf{e}_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right)$. Here, $\mathbf{V}_{i,11}^{(t)}$ is the element among the n_i diagonal elements of $\mathbf{V}_i^{(t)}$ corresponding to the reference sample. The rest of the elements $\mathbf{V}_{i,-1,-1}^{(t)}$ correspond to the other tumor samples. And the error covariance matrix for K outcomes in the i-th batch is $\boldsymbol{\Sigma}_i^{(t)} = \mathbf{Z}_i \boldsymbol{\tau}^{(t)} \boldsymbol{\tau}^{\prime(t)} \mathbf{Z}_i' + \mathbf{R}_i^{(t)} = \begin{pmatrix} \boldsymbol{\Sigma}_{i,oo}^{(t)} & \boldsymbol{\Sigma}_{i,om}^{(t)} \\ \boldsymbol{\Sigma}_{i,mo}^{(t)} & \boldsymbol{\Sigma}_{i,mm}^{(t)} \end{pmatrix}$.

In each M-step, we re-estimate the missing-data parameters (ϕ_0 and ϕ_1). Directly maximizing the likelihood function with the exponential missing-data mechanism in (2.3) may have convergence issues if the value of the exponential function (and the missing probability) exceeds one. Following Lumley *and others* (2006), we adopt a Poisson working model, instead of the exponential function as the missing-data mechanism for estimating ϕ_0 and ϕ_1. This would avoid the convergence issues and provide consistent estimates (Lumley *and others*, 2006).

To monitor the convergence, we derive the observed-data log-likelihood function as

$$
\begin{aligned}
l^o(\boldsymbol{\gamma}) = const &+ \sum_{i=1}^{N}\sum_{k\in\mathbf{O}_i}\log\left\{1-\exp\left(\phi_0+\phi_1/\mathrm{n_i}\cdot\mathbf{1}'\mathbf{y}_{\mathrm{ik}}\right)\right\} \\
&- \frac{1}{2}\sum_{i=1}^{N}\log\left|\boldsymbol{\Sigma}_{i,oo}\right| - \frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{y}_i^o-\mathbf{X}_i^o\boldsymbol{\alpha}\right)'\boldsymbol{\Sigma}_{i,oo}^{-1}\left(\mathbf{y}_i^o-\mathbf{X}_i^o\boldsymbol{\alpha}\right) \\
&+ \sum_{i=1}^{N}\left[K_i^m\phi_0+\phi_1/n_i\cdot\mathbf{1}'\left\{\mathrm{E}\left(\mathbf{y}_i^m|\mathbf{y}_i^o\right)+\phi_1/n_i\cdot\mathrm{var}\left(\mathbf{y}_i^m|\mathbf{y}_i^o\right)\mathbf{1}/2\right\}\right],
\end{aligned}
$$

where $\mathbf{O}_i$ denotes the set of indices for the observed outcomes in $\mathbf{Y}_i$, and K_i^m is the number of missing outcomes in the i-th batch. At the convergence, we derive the variance for fixed-effects estimates from the estimated information matrix of the observed-data log-likelihood function, $\widehat{\mathrm{var}}\left(\hat{\boldsymbol{\alpha}}\right) = \left(\sum_{i=1}^{N}\mathbf{X}_i^{o\prime}\hat{\boldsymbol{\Sigma}}_{i,oo}^{-1}\mathbf{X}_i^o\right)^{-1}$. One may construct the Wald statistics for testing fixed effects.

3.2. *A penalized EM-ADMM algorithm for the mvMISE$_e$ model*

In this section, we derive an EM algorithm for the mvMISE$_e$ model with correlated error terms. The E-step of the penalized EM algorithm was derived and is presented in Appendix B of the supplementary material available at *Biostatistics* online. The graphical lasso penalty on the precision matrix in (2.5) only affects the M-step. We employ an alternating direction method of multipliers (ADMM) algorithm (Boyd *and others*, 2011) for estimating the precision matrix in the M-step. Specifically, we first obtain the MLEs for the missing-data mechanism parameters (ϕ_0 and ϕ_1) as discussed in Section 3.1. We then obtain the estimates of the covariance matrix for the random effects

$$
\mathbf{D}^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N}\left\{\mathrm{E}\left(\mathbf{b}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\mathrm{E}\left(\mathbf{b}_i'|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)+\mathrm{var}\left(\mathbf{b}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)+\mathrm{var}\left(\mathbf{b}_i|\mathbf{y}_i,\boldsymbol{\gamma}^{(t)}\right)\right\},
$$

and the estimates of the fixed effects

$$
\boldsymbol{\alpha}^{(t+1)} = \left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{R}_i^{(t)^{-1}}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{N}\left[\mathbf{X}_i'\mathbf{R}_i^{(t)^{-1}}\left\{\mathrm{E}\left(\mathbf{y}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)-\mathbf{Z}_i\mathrm{E}\left(\mathbf{b}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\right\}\right],
$$

where $\boldsymbol{\gamma} = \{\boldsymbol{\alpha},\mathbf{D},\sigma_0^2,\sigma^2,\boldsymbol{\Sigma},\phi_0,\phi_1\}$ here.

The estimates of the error variances are obtained from the derivative of the log-likelihood function associated with $\mathbf{R}_i$ with respect to σ_0^2,σ^2 and $\boldsymbol{\Sigma}$,

$$
\begin{aligned}
l\left(\sigma_0^2,\sigma^2,\boldsymbol{\Sigma}\right) = const &- \frac{1}{2}\sum_{i=1}^{N}\log|\mathbf{R}_i| - \frac{1}{2}\sum_{i=1}^{N}\mathrm{E}\left(\mathbf{e}_i'|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\mathbf{R}_i^{-1}\mathrm{E}\left(\mathbf{e}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) \\
&- \frac{1}{2}\sum_{i=1}^{N}\mathrm{tr}\left\{\mathbf{R}_i^{-1}\mathrm{var}\left(\mathbf{e}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\right\}.
\end{aligned}
$$

Following Glanz and Carvalho (2018), we calculate the derivative of the third term above, the quadratic form of $\mathrm{E}\left(\mathbf{e}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)$. To obtain closed-form solutions for variance parameters, with Kronecker products ($\mathbf{R}_i = \boldsymbol{\Sigma}\otimes\mathbf{S}_i$) involved, we rewrite $\mathrm{var}\left(\mathbf{e}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)$ as the Kronecker product singular value

decomposition (Van Loan, 2000), var $\left(\mathbf{e}_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) = \sum_k \delta_{ik}\mathbf{G}_{ik} \otimes \mathbf{H}_{ik}$. Based on the properties of the Kronecker product, we derive the closed-form estimates:

$$\sigma_0^{2(t+1)} = \frac{1}{NK}\sum_{i=1}^{N}\tilde{\mathbf{V}}_{i,11}^{(t)}, \quad \text{and} \quad \sigma^{2(t+1)} = \frac{1}{K\left(\sum_{i=1}^{N} n_i - N\right)}\sum_{i=1}^{N}\text{tr}\left(\tilde{\mathbf{V}}_{i,-1,-1}^{(t)}\right),$$

where $\tilde{\mathbf{V}}_i^{(t)} = \tilde{\mathbf{E}}_i{\boldsymbol{\Sigma}^{(t)}}^{-1}\tilde{\mathbf{E}}_i' + \sum_k \delta_{ik}\text{tr}\left(\mathbf{G}_{ik}{\mathbf{S}_i^{(t)}}^{-1}\right)\mathbf{H}_{ik}'$, and $\tilde{\mathbf{E}}_i$ is an $n_i \times K$ matrix with $\text{vec}(\tilde{\mathbf{E}}_i) = \text{E}\left(\mathbf{e}_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right)$.

The major challenge in the mvMISE$_e$ model is the estimation of the error precision matrix, $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, when K is large. To estimate $\boldsymbol{\Theta}$ in each M-step, we first rewrite the log-likelihood function as a function of the error precision matrix:

$$l(\boldsymbol{\Theta}) = const + \frac{1}{2}\sum_{i=1}^{N} n_i \log|\boldsymbol{\Theta}| - \frac{1}{2}\sum_{i=1}^{N}\text{tr}\left(\tilde{\mathbf{E}}_i'\mathbf{S}_i^{-1}\tilde{\mathbf{E}}_i\boldsymbol{\Theta}\right) - \frac{1}{2}\sum_{i=1}^{N}\text{tr}\left\{\left(\boldsymbol{\Theta}\otimes\mathbf{S}_i^{-1}\right)\left(\sum_k \delta_{ik}\mathbf{G}_{ik}\otimes\mathbf{H}_{ik}\right)\right\}. \tag{3.1}$$

We then obtain the penalized MLEs for $\boldsymbol{\Theta}$ by maximizing the penalized log-likelihood in equation (2.5), which is the likelihood function in (3.1) plus a graphical lasso penalty on $\boldsymbol{\Theta}$.

Following Danaher *and others* (2014), we use the ADMM algorithm. The problem in (2.5) is equivalent to minimize$_{\boldsymbol{\Theta}}$ $-2l(\boldsymbol{\Theta}) + \lambda N|\mathbf{T}|_1$ subject to $\mathbf{T} = \boldsymbol{\Theta}$. With an additional penalty (the augmentation) and a Lagrange multiplier matrix $\mathbf{U}$ scaled by a penalty parameter ρ, we write the scaled augmented Lagrangian as

$$L_\rho(\boldsymbol{\Theta}, \mathbf{T}, \mathbf{U}) = -2l(\boldsymbol{\Theta}) + \lambda N|\mathbf{T}|_1 + \frac{\rho N}{2}\|\boldsymbol{\Theta} - \mathbf{T} + \mathbf{U}\|_F^2 - \frac{\rho N}{2}\|\mathbf{U}\|_F^2,$$

and here we use $\rho = 1$. The idea is to minimize $L_\rho(\boldsymbol{\Theta}, \mathbf{T}, \mathbf{U})$ with respect to $\boldsymbol{\Theta}$ and $\mathbf{T}$, respectively, and then update $\mathbf{U}$ at each iteration. Algorithm 1 provides the details for re-estimating the regularized $\boldsymbol{\Theta}$ within each iteration of the M-step. The tuning parameter (λ) can be selected by Akaike information criterion (Danaher *and others*, 2014).

Algorithm 1 The ADMM algorithm for re-estimating regularized $\boldsymbol{\Theta}$ within the M-step

1. Initialize with $\boldsymbol{\Theta} = \mathbf{I}, \mathbf{U} = \mathbf{T} = \mathbf{0}$.
2. Minimize the target function $-2l(\boldsymbol{\Theta}) + \frac{\rho N}{2}\|\boldsymbol{\Theta} - \mathbf{T} + \mathbf{U}\|_F^2$ with respect to $\boldsymbol{\Theta}$.
   Let $\Lambda\Omega\Lambda'$ be the eigendecomposition of the derivative of the target function, $\frac{1}{\sum_{i=1}^N n_i}\sum_{i=1}^N \mathbf{W}_i + \frac{\rho N\left(\mathbf{U}^{(t)} - \mathbf{T}^{(t)}\right)}{\sum_{i=1}^N n_i}$, where $\mathbf{W}_i = \tilde{\mathbf{E}}_i'{\mathbf{S}_i^{(t)}}^{-1}\tilde{\mathbf{E}}_i + \sum_k \delta_{ik}\text{tr}\left(\mathbf{H}_{ik}{\mathbf{S}_i^{(t)}}^{-1}\right)\mathbf{G}_{ik}'$. We have the estimate $\boldsymbol{\Theta}^{(t+1)}$ as $\Lambda\tilde{\Omega}\Lambda'$, where $\tilde{\Omega}$ is a diagonal matrix with the i-th diagonal element as $\frac{\sum_{i=1}^N n_i}{2\rho N}\left(-\omega_{ii} + \sqrt{\omega_{ii}^2 + \frac{4\rho N}{\sum_{i=1}^N n_i}}\right)$, where ω_{ii} is the i-th diagonal element of Ω.
3. Minimize $\lambda N|\mathbf{T}|_1 + \frac{\rho N}{2}\|\mathbf{T} - \boldsymbol{\Theta} - \mathbf{U}\|_F^2$ with respect to $\mathbf{T}$, where $|\mathbf{T}|_1 = \sum_{i\neq j}\left|\mathbf{T}_{ij}\right|$.
   Let $\mathbf{A} = \boldsymbol{\Theta} + \mathbf{U}$. We have $T_{ii}^{(t+1)} = A_{ii}^{(t)}, i = 1, \ldots, K$, for diagonal elements, and for $i \neq j$, $T_{ij}^{(t+1)} = \text{sgn}\left(A_{ij}^{(t)}\right)\left(\left|A_{ij}^{(t)}\right| - \lambda/\rho\right)_+$.
4. Update $\mathbf{U}^{(t+1)} = \mathbf{U}^{(t)} + \boldsymbol{\Theta}^{(t+1)} - \mathbf{T}^{(t+1)}$.
5. Iterate Step 2–4 until convergence.

Table 1. *Comparison of MSEs of the fixed effects estimates:* `mvMISE`$_b$ *versus* `lm` *and* `mixEMM` *when data have correlated random effects*

Parameters				MSE for intercept		MSE for slope		
Missingness	K	m_K	m_1	mixEMM	mvMISE$_b$	lm	mixEMM	mvMISE$_b$
exp	5	0.55	0.05	0.481	**0.060**	*0.067*	0.055	**0.044**
	20	0.55	0.00	0.066	**0.049**	*0.016*	0.012	**0.011**
	100	0.55	0.00	0.058	**0.041**	*0.003*	**0.002**	**0.002**
probit	5	0.54	0.05	0.386	**0.057**	*0.064*	0.053	**0.043**
	20	0.54	0.00	0.068	**0.049**	*0.014*	0.011	**0.010**
	100	0.54	0.00	0.063	**0.045**	*0.003*	**0.002**	**0.002**

In the "exp" and "probit" missingness settings, the incomplete data were generated based on the exponential function and the probit function, respectively, and had similar missing rate for K-variate outcomes (m_K) and for averaged univariate outcome (m_1). The model estimation is based on the exponential function. We set $\sigma_0^2 = 1$ (experimental variation for the reference sample), $\sigma^2 = 2$ (biological variance), $\phi_1 = -0.015$ (missing data parameter, a non-zero value implies non-ignorable abundance-dependent missing data). We simulated a binary predictor of interest. We set $\alpha_0 = 10$ (the intercept), and the slope $\alpha_1 = 1.3/\sqrt{K}$. We specified the variance components of factor-analytic random effect $\boldsymbol{\tau}$ to have elements $\tau = \sqrt{7U/3}$, where $U \sim \text{Unif}\left[\sigma_0^2 - 1, \sigma_0^2 + 1\right]$. The results were based on 1000 replications. The smaller/smallest MSE in each setting is shown in boldface.

4. Simulations

4.1. *Comparing the mean squared errors (MSEs) and biases of fixed effects estimates*

We conducted simulations to assess the MSEs and biases of the proposed methods versus competing methods. We compared the following methods: (1) the standard practice in the current proteomics literature using linear regression (`lm`) on the relative abundance—the abundance of a protein/peptide in the observed sample relative to the reference sample in the same batch; (2) the univariate mixed-effects model (`mixEMM`) by Chen *and others* (2017b); and (3) the proposed multivariate selection models (`mvMISE`$_b$ or `mvMISE`$_e$). Note that methods (1) and (2) are univariate analysis methods, and when applying those methods to analyze each protein, we took the average abundances of peptides from each protein as the analysis unit. Also note that method (1) is based on relative abundances and as such would not offer the mean estimates for absolute protein abundance levels. Methods (2) and (3) are based on absolute abundances. In each simulation, we simulated a total of 108 samples from 36 clusters/batches, where each batch consists of four samples including a common reference sample. This is the same sample size as in the CPTAC data.

To compare the mvMISE$_b$ method with competing methods, we simulated multivariate outcome ($K =$ 5, 20, and 100) with incomplete data based on the models (2.3) and (2.4). To assess the robustness of the proposed methods to misspecification of the missing-data function, we also simulated incomplete data based on a probit function in (2.2) and applied mvMISE$_b$ based on the exponential function. In each simulation setting, we simulated a fixed intercept (α_0) and fixed effects of interest (α_1). Herein, we simulated a common α_1 for different peptides in the same protein. We repeated the simulations 1000 times. We set the simulation parameters to mimic the data structure that has been observed in the CPTAC data. Detailed simulation parameters were provided in the caption of Table 1. Table 1 shows that the mvMISE$_b$ method has consistently smaller MSEs than mixEMM for both intercept (α_0) and slope (α_1) in different settings, especially when K is small (i.e., 5). The linear regression based on relative abundance does not offer comparable mean estimates of absolute protein abundances and is omitted in the comparison for intercept. Its slope estimates are also not directly comparable, but one can see that when K is small, the method tends to have large MSEs. This finding is shared with all methods. The bias comparison of competing methods reached similar conclusions and is presented in Table S1 in the supplementary material

Table 2. *Comparison of MSEs of the fixed effects estimates:* `mvMISE`$_e$ *versus* `lm` *and* `mixEMM` *when data have correlated error terms (in protein pathway analysis)*

Parameters				MSE for intercept		MSE for slope		
Missingness	K	m_K	m_1	mixEMM	mvMISE$_e$	lm	mixEMM	mvMISE$_e$
exp	5	0.55	0.05	0.547	**0.039**	*0.068*	0.053	**0.040**
	20	0.55	0	**0.029**	**0.029**	*0.016*	0.013	**0.012**
	100	0.55	0	**0.027**	**0.027**	*0.003*	**0.002**	**0.002**
probit	5	0.65	0.12	2.330	**0.051**	*0.092*	0.075	**0.058**
	20	0.65	0	0.032	**0.031**	*0.022*	0.018	**0.016**
	100	0.65	0	**0.027**	**0.027**	*0.004*	**0.003**	**0.003**

In the "exp" and "probit" missingness settings, the incomplete data were generated based on the exponential function and the probit function, respectively, and had similar missing rate for K-variate outcomes (m_K) and for averaged univariate outcome (m_1). The model estimation is based on the exponential function. We set $\sigma_0^2 = 1$ (experimental variation for the reference sample), $\sigma^2 = 2$ (biological variance), $\phi_1 = -0.015$ (missing data parameter, a non-zero value implies non-ignorable abundance-dependent missing data). We simulated a binary predictor of interest. We set $\alpha_0 = 10$ (the intercept), and the slope $\alpha_1 = 1.4/\sqrt{K}$. The results were based on 1000 replications. The smaller/smallest MSE in each setting is shown in boldface.

available at *Biostatistics* online. This simulation also shows that the estimation and inference are robust to misspecification of the true missing-data model.

To compare the mvMISE$_e$ method with competing methods, we simulated multivariate correlated outcomes ($K = 5$, 20, and 100) with a sparse inverse covariance matrix, mimicking the correlations among multiple proteins in a pathway. For each simulation setting, we simulated 1000 incomplete data based on the exponential missing-data function and 1000 data based on the probit missing-data function. In each setting, we simulated K proteins based on model (2.1) with a random intercept and error term $\mathbf{e}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{S}_i)$, where $\boldsymbol{\Sigma}$ captures the biological correlations among proteins. Following Danaher *and others* (2014), we simulated a sparse biological correlation structure from a power-law degree distribution. Other simulation parameters are similar as in the simulation for mvMISE$_b$ (see Table 2 caption for details).

We simulated a fixed intercept (α_0) and a fixed effect of interest (α_1) that is common to all proteins within a pathway. Note that the fixed effects for different proteins in a pathway need not be the same. This specification is only for the ease of comparing biases and MSEs of different methods in the current setting. Table 2 shows that the proposed mvMISE$_e$ model has smaller MSEs than mixEMM in most, if not all, settings. Comparison of biases reached similar conclusions and is presented in Table S2 of the supplementary material available at *Biostatistics* online. The mvMISE$_e$ model estimation is also robust to misspecification of the missing-data function.

4.2. *Comparing the type I error rates and power in testing for fixed effects*

4.2.1. *The mvMISE$_b$ model.* To compare the type I error rates and power for mvMISE$_b$ versus competing methods, we simulated the data similarly as in Section 4.1. When applying the mvMISE$_b$ model to the abundances of K simulated peptides, we obtained the fixed effects estimates, calculated the standard errors using the observed information matrix from the EM algorithm, and obtained the P-values based on the Wald Z-test (see Section 3.1 for details).

Table 3 shows that the type I error rates for our mvMISE$_b$ model with parametric P-values are well controlled at the nominal level (0.05). However, the linear regression based on relative abundances has a deflated type I error rate in most cases, resulting in a conservative test. The univariate mixEMM model ignores the correlations among multiple peptides and has slightly inflated type I error rates in some settings.

Table 3. *Comparison of type I error rates and power for testing non-zero fixed effects* α_1*:* `mvMISE`$_b$ *versus* `lm` *and* `mixEMM`*, when data were generated with correlated random effects*

	Missingness	exp			probit		
	K	5	20	100	5	20	100
Type I error rate	lm	0.029*	0.034*	0.050	0.027*	0.021*	0.040
	mixEMM	0.060	0.050	0.069*	0.050	0.050	0.050
	mvMISE$_b$	0.052	0.051	0.056	0.051	0.046	0.055
Power	lm	0.534	0.610	0.602	0.551	0.601	0.600
	mixEMM	0.739	0.780	**0.837***	0.762	0.787	**0.810**
	mvMISE$_b$	**0.807**	**0.803**	0.805	**0.816**	**0.797**	0.797

The simulation parameters are the same as those in Table 1 except $\alpha_1 = 0$ (under the null) when assessing the type I error rate. The results were based on 1000 replications in each setting. Type I error rates significantly differ from the nominal level 0.05 based on a binomial test are marked with a star, i.e., type I error rate higher than or equal to 0.064, or lower than or equal to 0.036. The highest power in each setting is shown in boldface.

Table 3 also shows that the proposed method mvMISE$_b$ improved the power for testing non-zero fixed effects as compared to the univariate methods in most of the settings in which the type I error rates can be properly controlled.

4.2.2. *The mvMISE$_e$ model.* To assess the type I error rates and power for the mvMISE$_e$ model in pathway analyses of multiple proteins, we simulated the data similarly as in the Section 4.1. When comparing power, here, we simulated each protein in a pathway to have a protein-specific effect (see Table 4 caption for details). All of the results were based on 1000 replications. In the analyses of the univariate methods, `lm` and `mixEMM`, we first took the average peptide abundances as the protein abundances, estimated the protein-specific effects using the corresponding methods, and calculated the nominal protein-level P-values based on the Wald tests. We then combined those protein-level P-values in a pathway using Fisher's method. The P-values for each pathway are calculated based on 1000 permutations. In the mvMISE$_e$ analysis, we took the average peptide abundances as the protein abundances while jointly considering multiple proteins in a pathway. We simultaneously estimated the Wald statistic for each protein in the pathway in the mvMISE$_e$ model and calculated the P-values. We then combined those protein-level P-values in the pathway using Fisher's method and conducted permutations. Note that here, we used Fisher's method as the pathway analysis method to aggregate protein-level P-values within pathways. One may use a different gene-set analysis approach, but it does not affect the conclusions of our comparisons.

To assess the type I error rates, we set all protein-specific effects to be zero. We calculated the permutation-based P-values for all methods. Table 4 shows that, with permutations, the type I error rates for `lm`, `mixEMM` and the proposed `mvMISE`$_e$ models were almost all well-controlled at the nominal level (0.05). For power comparison, we simulated four different types of alternatives (see Table 4 caption for details). The highest power in each setting is shown in boldface in Table 4. The proposed mvMISE$_e$ model is most powerful in all settings.

5. Applications to the CPTAC breast cancer data

The CPTAC (http://proteomics.cancer.gov) is a comprehensive and co-ordinated effort launched by the National Cancer Institute (Ellis *and others*, 2013). The overall goal of CPTAC is to improve our ability

Table 4. *Comparison of type I error rates and power for testing non-zero fixed effects α_1: Fisher's method-based pathway analysis using P-values from* `mvMISE`$_e$ *versus* `lm` *and* `mixEMM`*, when data were generated with potentially correlated error terms*

		Missingness	exp			probit		
	Setting	K	5	20	100	5	20	100
Type I error rate	Under the null (correlated errors)	lm	0.065*	0.059	0.055	0.049	0.052	0.049
		mixEMM	0.045	0.059	0.038	0.059	0.069*	0.050
		mvMISE$_e$	0.062	0.054	0.053	0.054	0.057	0.054
Power	All equal signals $\alpha_k = 0.6/\log_{10}(K)$ (correlated errors)	lm	0.772	0.592	0.607	0.612	0.430	0.398
		mixEMM	0.856	0.672	0.730	0.653	0.499	0.486
		mvMISE$_e$	**0.913**	**0.793**	**0.907**	**0.766**	**0.677**	**0.751**
	All equal signals $\alpha_k = 0.6/\log_{10}(K)$ (independent errors)	lm	0.776	0.602	0.636	0.622	0.440	0.431
		mixEMM	0.850	0.660	0.739	0.651	0.480	0.540
		mvMISE$_e$	**0.907**	**0.804**	**0.905**	**0.745**	**0.652**	**0.770**
	Half signals $\alpha_k = 0.8/\log_{10}(K)$ (correlated errors)	lm	0.702	0.499	0.589	0.576	0.320	0.361
		mixEMM	0.741	0.552	0.671	0.643	0.396	0.440
		mvMISE$_e$	**0.793**	**0.691**	**0.828**	**0.693**	**0.526**	**0.658**
	Random signals (correlated errors)	lm	0.724	0.547	0.592	0.601	0.380	0.385
		mixEMM	0.780	0.557	0.655	0.614	0.416	0.430
		mvMISE$_e$	**0.803**	**0.701**	**0.824**	**0.688**	**0.585**	**0.675**

The simulation parameters are similar to those in Table 2. We set $\alpha_k = 0$ when assessing the type I error rates. When comparing power, we simulated four different types of alternatives. In the "All equal signals" setting, we set $\alpha_k = 0.6/\log_{10}(K)$ for all outcomes, allowing the error terms to be correlated or independent; in the "Half signals" setting, we set half of the $\alpha_k = 0.8/\log_{10}(K)$ and the rest to zero; and in the "Random signals" setting, the outcome-specific effect sizes are randomly generated from Unif$(-1/\log_{10}(K), 1/\log_{10}(K))$. Note that all the P-values are calculated based on permutation.
The results were based on 1000 replications in each setting. Type I error rates significantly differ from the nominal level 0.05 based on a binomial test are marked with a star, i.e., type I error rate higher than or equal to 0.064, or lower than or equal to 0.036. The highest power in each setting is shown in boldface.

to diagnose, treat and prevent cancer through the application of robust, quantitative, proteomic technologies and workflow. The consortium has recently conducted global proteome and phosphoproteome profiling of many breast, colon and ovarian cancer samples. In this work, we focused on analyzing the phosphoproteomics data from the CPTAC breast cancer study (Mertins *and others*, 2016). Phosphorylation is a key post-translational modification and plays major roles in many biological processes. Different phosphorylation sites (phosphopeptides) of one protein could induce different biological activities. Meanwhile, the phosphopeptides from the same phosphoprotein could be highly correlated in abundances.

In the CPTAC breast cancer dataset, a total of 108 tumor samples from 104 women with breast cancer and aged 26–90 were randomly assigned to 36 batches and were processed by 36 four-plex (i.e., four-channel) iTRAQ experiments. In each batch of samples, there were three tumor samples and one common reference sample created by combining 40 tumor samples. In the following analyses, we focused on 77 tumor samples with superior quality. After standard data preprocessing, we analyzed the abundance data for 25 961 phosphopeptides that were observed in at least 25 (70%) of the 36 runs of the reference sample. Those phosphopeptides correspond to 6078 phosphoproteins. Further details of the data are in Appendix C of supplementary material available at *Biostatistics* online. Figure 1(a) shows the histogram of the number of phosphopeptides for each phosphoprotein. Most phosphoproteins have no more than five phosphopeptides, though some can have nearly 200 phosphopeptides. Figure 1(b) showed the distribution

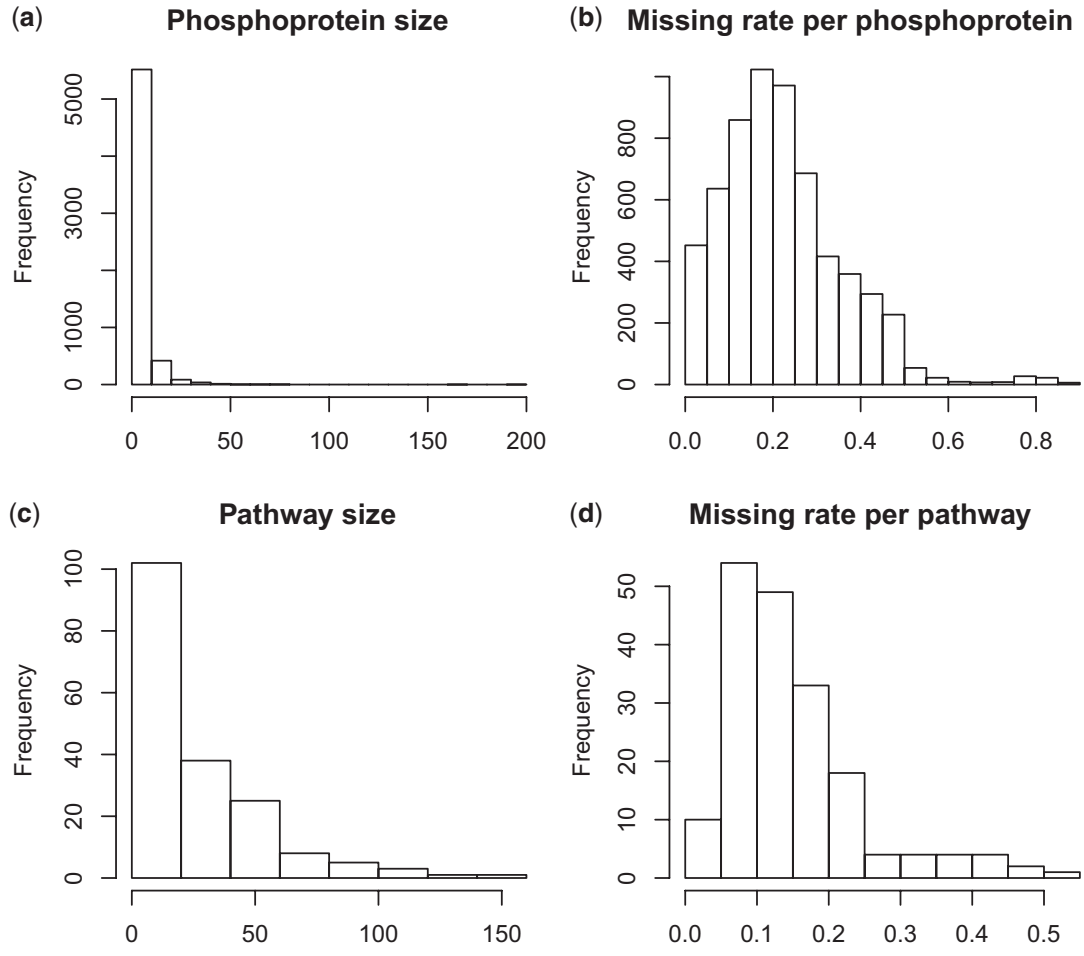

Fig. 1. A summary of the CPTAC breast cancer data. (a) The histogram of the number of phosphopeptides in each phosphoprotein. There were two very large phosphoproteins with 164 and 195 phosphopeptides. (b) The histogram of the average missing rate for each phosphoprotein averaged across phosphopeptides. The mean and median of average missing rates were 22.2% and 20.3%, respectively. (c) The histogram of the number of proteins in each KEGG pathway. (d) The histogram of the average missing rate in each KEGG pathway averaged across protein. The mean and median of average missing rates were 13.7% and 11.1%, respectively.

of average missing rate for phosphoprotein averaged across phosphopeptides. The median of average missing rate is 20.3%.

5.1. *Model checking*

Given the substantial amount of missing data and the demonstrated non-ignorable abundance-dependent missing-data mechanism, in the mvMISE methods we model the batch-level missing-data probability as in equation (2.3) and incorporate it in the likelihood. In this data set, we observed a strong linear relationship between the log of the percentage of missing batches for each peptide versus the mean peptide abundance based on the observed peptides (see Figure 3 for a graphical illustration in Chen *and others* (2017b)).

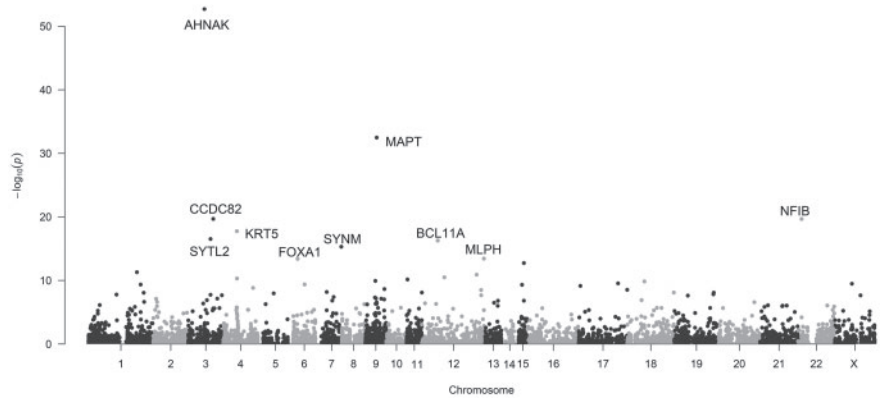

Fig. 2. The Manhattan plot of the phosphoproteins identified by the proposed mvMISE$_b$ method. The names of the top ten most significant proteins were labeled.

Therefore, we model the batch-level missing probability as an exponential function of the average peptide abundances.

When applying the proposed models to other abundance-dependent incomplete data, we recommend checking the fit of the selected and estimated missing-data mechanism function before applying the mvMISE methods in the estimation and inference. If the function in (2.3) provides a poor fit to real dataset, one may consider employing the more flexible probit function or other functions as the link function in model (2.2) and incorporate the fitted missing-data model in the likelihood-based estimation and inference.

5.2. *Analysis I: using mvMISE$_b$ to jointly analyze multiple phosphopeptides of a phosphoprotein*

For each phosphoprotein, we treated the abundance levels of multiple phosphopeptides of this phosphoprotein as the multivariate response variable. We used an indicator for triple negative breast cancer (TNBC) tumor sample as the predictor of interest and also included an indicator for the reference sample. We applied the proposed mvMISE$_b$ method to each of the 6078 phosphoproteins. We obtained P-values based on the parametric Wald tests for non-zero effects of the TNBC indicator.

Figure 2 shows the Manhattan plot of the mvMISE$_b$-based P-values. At the Bonferroni-adjusted P-value threshold of 8.23×10^{-6}, there were 119 out of 6078 phosphoproteins significantly associated with TNBC subtype. Among the 119 significant proteins, 77 were uniquely identified by our method and were not identified by the standard practice of linear regression based on relative abundance protein measures using 10 000 permutations. This showed our method as a useful complementary approach to the standard practice. Our top protein is AHNAK, with a P-value of 1.93×10^{-53}. The gene *AHNAK* negatively regulates cell growth and acts as a tumor suppressor of the TGF-β signaling pathway. More recently, Chen *and others* (2017a) also suggested that AHNAK suppresses tumor proliferation and invasion by targeting multiple pathways, including the MAPK signaling pathway, in TNBC patients. In our data, the protein AHNAK has 195 phosphopeptides, and about half of them were individually significant at the P-value threshold of 0.05 when analyzing each phosphopeptide by `mixEMM`. This further illustrated the advantage of our methods over simple averaging when analyzing large proteins with many phosphopeptides.

5.3. *Analysis II: using mvMISE$_e$ in protein pathway analyses*

We applied the mvMISE$_e$ method to identify KEGG (Kyoto Encyclopedia of Genes and Genomes) human disease pathways with proteins showing different activities between TNBC tumors versus other

Table 5. *The significant KEGG pathways detected by the proposed mvMISE$_e$ method with a Bonferroni-adjusted P-value cut-off of 0.1 (i.e., P-value $\leq 0.1/150 = 6.67\times10^{-4}$)*

KEGG human disease pathway name	No. of proteins	No. (%) proteins marginally significant	mvMESM$_e$ *P*-values	Marginally significant proteins in the pathway
MAPK signaling pathway	130	18 (13.85)	0.0006	DUSP9, EGFR, FGFR3 FLNB, HSPA8, MAP2K4 MAP2K5, MAP3K1, MAP3K11 MAP3K4, MAPK11, MAPK8IP3 MAPT, MYC, NLK PDGFRB, PRKCA, TP53
Non-small cell lung cancer	34	9 (26.47)	0.0005	AKT2, EGFR, ERBB2 FOXO3, PIK3R1, PIK3R5 PLCG2, PRKCA, RXRA
Inositol phosphate metabolism	31	8 (25.81)	0.0004	IMPA1, INPP4A, INPP5E INPP5J, PIK3CG, PIP4K2B PLCG2, SYNJ1
Thyroid cancer	17	3 (17.65)	0.0006	CTNNB1, MYC, RXRA

The results were based on 10 000 permutations.

breast cancer tumors. Among the 186 KEGG pathways, there were 150 pathways that have at least five phosphoproteins being mapped in the CPTAC data. Figure 1(c) shows the histogram of the number of phosphoproteins in each mapped KEGG pathway. Figure 1(d) shows the average missing rate for phosphoproteins in each pathway averaged across proteins.

We first obtained the protein-level abundances as described earlier. After standardizing protein abundances and treating them as multivariate responses in mvMISE$_e$, we estimated the protein-specific effects on TNBC subtype by introducing the interactions between the protein indicators and the TNBC indicator and obtaining the protein-specific Wald statistics. For a pathway with K phosphoproteins, we estimated one intercept, one common effect of the reference sample, and K protein-specific TNBC effects. Due to the regularization of parameter estimation, we calculated the P-value for each protein, used Fisher's method to aggregate protein-level P-values to a pathway, and then obtained the P-value for each pathway based on 10 000 permutations.

Table 5 lists the significant KEGG pathways detected by the mvMISE$_e$ method at the Bonferroni-adjusted P-value cut-off of $0.1/150 = 6.67\times10^{-4}$. Also listed are the number of phosphoproteins in each significant pathway and the phosphoproteins that are individually significantly associated with TNBC at the level of 0.05. One can see that all of the significant findings of mvMISE$_e$ are driven by individually weak-to-moderate and collectively strong associations to TNBC. Among those pathways, the MAPK signaling pathway is known to be related to breast cancer risk and in particular TNBC (Giltnane and Balko, 2014). Eighteen out of 130 proteins in this pathway showed marginally significant evidence of associations to TNBC. A known cancer gene, *MYC*, is marginally significant and is in both the MAPK signaling and the thyroid cancer pathway. Some other genes (*EGFR* and *PRKCA*) shared by the MAPK signaling and non-small cell lung cancer pathways have also been shown to be related to TNBC risk.

6. Discussion

Motivated by the batch-processed CPTAC proteomics data, in this work, we developed two tailored multivariate mixed-effects selection models for simultaneously analyzing multiple peptides in a protein or

multiple proteins in a pathway with clustered data structure and non-ignorable missing data. The proposed models complement the univariate approach proposed in Chen *and others* (2017b) and aim to address the challenges arising from batch-processed proteomics data analysis, including the clustered data structure, the batch-level non-ignorable missing data mechanism, and the correlation structures for high-dimensional outcomes.

Although motivated by the multivariate analysis of proteomics data, the proposed models and methods are flexible and generalizable to other data types of similar characteristics and structures. Those models can also be used in analyzing high-dimensional outcomes with clustered data and other correlation structures among high-dimensional outcomes. Single or multiple imputations can also be performed based on extensions of the proposed models. Those imputations would facilitate other downstream analysis such as network analyses, in particular addressing the limitation that much of the standard software for such analyses cannot handle incomplete data as input data.

Before applying the proposed mvMISE models to other data sets, one needs to check the fit of the current exponential missing-data mechanism by examining whether there is a strong linear relationship of the log of missing rates versus the mean values of the individual response variables. Alternatively, the probit missing-data function or other missing-data mechanism functions can be integrated into the likelihood function to analyze clustered data with other ignorable or non-ignorable missing data mechanisms. After obtaining the results by mvMISE, one can always check whether the significant findings are reflecting individually weak-to-moderate and collectively strong effects on the set of response variables.

One caveat of the $\texttt{mvMISE}_e$-based pathway analysis is that we ignored the correlations among phosphopeptides in each protein and took the average phosphopeptide abundances as the protein-level abundance measures; we then applied the $\texttt{mvMISE}_e$ model to multiple protein abundances in a pathway. We used permutations to calculate *P*-values in pathway analyses, so this is less of an issue for significance control. An alternative approach is to directly model the hierarchical structures of multiple phosphopeptides nested in each protein and multiple proteins nested in a pathway. The development of this hierarchical model requires the joint modeling of correlated random effects among multiple phosphopeptides from each protein and correlated error terms among multiple proteins in a pathway. It is challenging to estimate such a large number of parameters when the sample size is limited. Some flexible Bayesian models may serve this purpose and those will be explored in future work.

We applied the methods to the CPTAC breast cancer proteomic dataset. We identified significant phosphoproteins that are related to TNBC risk and would otherwise not be detected by competing methods. We also identified pathways enriched with phosphoproteins with differential abundances in TNBC tumors versus other tumors. The proposed multivariate methods can serve as complementary methods to the univariate approaches, and the analyses may provide additional insights into breast cancer and subtype etiology such as identifying new protein biomarkers.

In other studies beyond labeled proteomic studies, there may be little information regarding the missing-data mechanism. One may also consider multivariate mixed-effects models for clustered data with other types of non-ignorable missingness, such as shared-parameter models, pattern-mixture models and mixed-effects hybrid models, for model comparison and sensitivity analysis. The current work may also be extended to general multivariate analyses for clustered data with clustered outcome-dependent missingness beyond quantitative proteomic studies.

We developed an R software package mvMISE. It is currently available at GitHub (https://github.com/randel/mvMISE) and will be made available through R CRAN.

SUPPLEMENTARY MATERIAL

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

Acknowledgments

Mass spectrometry and proteomics data were acquired by the breast cancer project of CPTAC consortium, which is led by Dr. Steve Carr from Broad Institute of MIT and Harvard, and Dr. Amenda Paulovich from Fred Hutchinson Cancer Research Center and is supported by NCI grant CA160034. We thank Dr. Chenwei Lin, Dr. D.R. Mani, Dr. Philipp Mertins, Dr. Yan Ping, and others from the CPTAC consortium for their help on the proteomics data. *Conflict of Interest*: None declared.

Funding

This work was supported by the National Institutes of Health [R01GM108711 to L.S.C., U24CA210993 to P.W. and L.S.C, SUB-CA160034 to P.W.].

References

Baladandayuthapani, V., Talluri, R., Ji, Y., Coombes, K. R., Lu, Y., Hennessy, B. T., Davies, M. A. and Mallick, B. K. (2014). Bayesian sparse graphical models for classification with application to protein expression data. *The Annals of Applied Statistics* **8**, 1443.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* **3**, 1–122.

Chen, B., Wang, J., Dai, D., Zhou, Q., Guo, X., Tian, Z. and Xie, X. (2017a). AHNAK suppresses tumour proliferation and invasion by targeting multiple pathways in triple-negative breast cancer. *Journal of Experimental & Clinical Cancer Research* **36**, 65.

Chen, L. S., Wang, J., Wang, X. and Wang, P. (2017b). A mixed-effects model for incomplete data from labeling-based quantitative proteomics experiments. *The Annals of Applied Statistics* **11**, 114–138.

Clough, T., Key, M., Ott, I., Ragg, S., Schadow, G. and Vitek, O. (2009). Protein quantification in label-free LC-MS experiments. *Journal of Proteome Research* **8**, 5275–5284.

Danaher, P., Wang, P. and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 373–397.

Ellis, M., Gillette, M., Carr, S., Paulovich, A., Smith, R., Rodland, K., Townsend, R., Kinsinger, C., Mesri, M., Rodriguez, H., Liebler, D. *and others*. (2013). Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discovery* **3**, 1108–1112.

Giltnane, J. M. and Balko, J. M. (2014). Rationale for targeting the Ras/MAPK pathway in triple-negative breast cancer. *Discovery Medicine* **17**, 275–283.

Glanz, H. and Carvalho, L. (2018). An expectation–maximization algorithm for the matrix normal distribution with an application in remote sensing. *Journal of Multivariate Analysis* **167**, 31–48.

Karp, N. A., Huber, W., Sadowski, P. G., Charles, P. D., Hester, S. V. and Lilley, K. S. (2010). Addressing accuracy and precision issues in iTRAQ quantitation. *Molecular & Cellular Proteomics* **9**, 1885–1897.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. New York: Wiley.

Liu, L. C. and Hedeker, D. (2006). A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics* **62**, 261–268.

Lumley, T., Kronmal, R. and Ma, S. (2006). Relative risk regression in medical research: models, contrasts, estimators and algorithms. University of Washington Biostatistics Working Paper Series, Working Paper 293. http://www.bepress.com/uwbiostat/paper293.

MERTINS, P., MANI, D. R., RUGGLES, K. V., GILLETTE, M. A., CLAUSER, K. R., WANG, P. scemphAND OTHERS. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55–62.

ROY, J. AND LIN, X. (2002). Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates: changes in methadone treatment practices. *Journal of the American Statistical Association* **97**, 40–52.

VAN LOAN, C. F. (2000). The ubiquitous Kronecker product. *Journal of Computational and Applied Mathematics* **123**, 85–100.

WIESE, S., REIDEGELD, K. A., MEYER, H. E. AND WARSCHEID, B. (2007). Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. *Proteomics* **7**, 340–350.

[*Received August 14, 2017; revised March 16, 2018; accepted for publication May 21, 2018*]