Statistics in Medicine
WILEY

A Bayesian approach to multivariate and multilevel modelling with non-random missingness for hierarchical clinical proteomics data

SCHOLARONE™
Manuscripts

A Bayesian approach to multivariate and multilevel modelling with non-random missingness for hierarchical clinical proteomics data

Irene SL Zeng[1] Thomas Lumley[1], Katya Ruggiero1[1], Martin Middleditch[2]

1. Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
2. Centre for Genomics and Proteomics, University of Auckland, New Zealand

Corresponding author:

Irene SL Zeng

Email: i.zeng@auckland.ac.nz

Telephone: 00649-3003977

Abstract

High throughput mass-spectrometry-based proteomics data from clinical studies brings challenges to statistical analysis. The challenges originate from the hierarchical levels of protein abundance data and interactions between clinical study design and experimental design. The non-random missingness of the measurements from a vast amount of information also adds complexity in data analysis. We propose multivariate multilevel models to analyse protein abundances and to handle abundance-dependent missingness within a Bayesian framework. The proposed model enables the variance decomposition at different levels of the data hierarchy and provides shrinkage of protein-level estimates for a group of proteins. A logistic missingness and censored model with informative prior is used to handle incomplete data. Hamiltonian MC/No-U-Turn Sampling and Gibb MCMC algorithms are created to derive the posterior distribution of study parameters; Hamiltonian MC is demonstrated to gain more efficiency for these high-dimensional correlated data.

Improvements of the proposed missing data model is compared to the univariate mixed effect model and the multivariate-multilevel model using complete data in a simulated study and a clinical proteomics study.   The proposed model framework can be used in other types of data with similar structure and Non Random Missingness mechanism (MNAR).

Keywords: Hamiltonian Monte Carlo; designed clinical proteomics experiment; multilevel models; Bayesian approach for non-random missingness.

1. Introduction

1.1. Clinical Proteomic studies

Proteomics has been used as a system biology technique in pharmaceutical industry and medical research to investigate hundreds and thousands of molecular biomarkers simultaneously. It belongs to the family of system biology, including metabolomics, transcriptomics and genomics. These "omic" fields study interacting molecular networks inter-disciplinarily. Like aforementioned "omics", proteomics utilizes biotechnology platforms that can systematically identify proteins and quantify their abundances. One of the popular platforms used for protein discovery is mass spectrometry, which can accurately determine the molecular mass of ions for peptide/protein sequencing and quantitation [1-4]. Through different ionization and detection processes, mass spectrometry (MS) coupled with other techniques (i.e. chromatography) can separate different peptide species from a biological sample and produce a large amount of ion intensity data [5-9]. These intensity data are used to inform the abundances of the peptides and from these to construct the abundances of the corresponding proteins. The *hierarchical structure* of the protein abundance originates from the biochemical process of the mass spectrometry-based proteomic experiment [10]. When a proteomic experiment involves more than one biological sample, through assay batching or multiplexing [11-14], multiple biological samples can be analysed simultaneously in a single-run experiment. A multiplex comprises a

batch of multiple samples. Each sample is labelled in the multiplex for its identity in a mass spectrometry experiment. When multiple runs are involved, the biological samples will be allocated into *labels* and *runs* according to an experimental design.

The statistical analysis of mass spectrometry (MS) data curated from multiple-run experiments of multiplex-assays requires us to understand the data structure fully. Firstly, due to biochemical processes (trypsin digestion and protein fragmentation), the abundance data is provided at the peptides level, not proteins. The proteins abundances must be derived from their observed constituent peptides' abundances. Secondly, each protein is not identified by the same set of peptides in each experiment. Thirdly, the MS experimental factors could have different impacts across different peptides/proteins. Fourthly, a clinical proteomics study has two phases of study design: clinical study design and biological experimental design. The confounding effects from a MS-based experiment hence are multileveled. Using a multilevel model will be an advanced solution to coping these confounding variables.

1.2. Data structure and sources of variation of a Mass Spectrometry-based isotopic labelling experiment

The protein expression data curated from MS-based proteomics experiments can either be the relative or absolute abundance of proteins. The relative abundances of proteins are measured by the intensity of ions from the mass spectrometry, while the absolute abundance of the protein is measured by the concentration [10]. The relative quantitation of protein abundance utilize different labelling techniques including chemical, biological, metabolic and enzymatic incorporation in a multi-sample proteomics experiment [10-12]. One common labelling experiment is to use iTRAQ, an isobaric chemical labelling approach that is widely adopted in proteomics laboratories. iTRAQ is a biochemical reagent comprising a reactive group, a tag (label) with a mass balance group and a reporter group [10-12, 14-18]. The reporter groups of the iTRAQ reagent, which have different molecular mass-to-charge ratios ranging from 114.1-117.1 Th in the 4-plex assay and 113.1-121.1 Th in the 8-plex assay, are used to distinguish multiple samples in the assay. The intensities of the labelled peptides are used to inform the relative abundance of proteins [19].

1.3. The instrumental feature of the mass spectrometer and the quantification

In addition to the hierarchical structure of protein abundance data, the physical features of mass spectrometry also influence the variations in ions intensities. We discovered from the two proteomic iTRAQ studies that, the logarithmic intensity decreases with the mass-to-charge ratio and the molecular mass of the peptides. The identified variations in intensity measurements caused by MS's physical natures in our case studies are consistent with the findings of the other studies [20, 21]. Breitwieser et al. [20] observed more variabilities for the lower intensity signals and illuminated a funnel-shaped association between the peptide ratio and the logarithmic intensities. This evidence suggested the relative abundance of a protein is likely to be underestimated by its large peptide components. The quantification of a protein thus needs to take into account the variations from instruments. The proposed multilevel model includes mass-to-charge ratios as a variable contributing to this variation and is shown to increase the precision of estimates for the unknown parameters.

1.4. The non-random missingness mechanism in data from mass spectromety

The missingness mechanisms for intensities data comprises both Missing-Not-At-Random (MNAR) and censoring. The MNAR mechanism of protein abundance is related to their peptides intensities, masses, electrical charges, and ionization efficiency [21, 22]. The likelihood of missing values thus may depend on the unknown protein abundance and the observable mass-to-charge ratio. The left-censored phenomenon is caused by the detecting limit of a MS. We model the probability of having a missing intensity as a function of the intensity and the mass-to-charge ratio [figure 1]. Data from mass spectrometry of smaller molecules (metabolomics) supports this model [21]. Their data showed that higher probabilities of having missing intensities were associated with observed lower abundance. There is a nonlinear relationship between probability of having noisy intensity value and the mass-to-charge ratio.

1.5. An imputation model under a Bayesian multivariate and multilevel inference framework

We use an inference model within a multivariate-multilevel framework to analyze the hierarchical protein abundance data generated by clinical proteomics studies. These

models decompose variations from different sources, such as experimental factors, biological features of proteins and physiological characteristics of biological samples. We further model non-random missingness as either left-censored or completely missing. Little [23], Little and Rubin [24] proposed three likelihood-based methods (selection model, pattern-mixture model and shared-parameter model) for MNAR. Selection model, in which the missing data mechanism and parameters of the inferential model are conditionally estimated given the hypothetical complete data, are the appropriate approach for the abundance-dependent missing data. The hypothetical complete data comprises the missing and observed values. The known physical properties of peptides and mass spectrometers provide the **auxiliary information** for the missing data. We use the selection model factorization [24] for MNAR including the left-censored missingness under a Bayesian framework. We estimate the likelihood of missingness by using auxiliary variable mass-to-charge ratio (m/z) and the intensity values. The censored and non-censored missing intensities are marginalized for the complete data inferential likelihood, i.e.

$$f(y|X) = \int \int f\left(y^{obs}|X^{obs}, B\right) \times f\left(y^{miss}|X^{miss}, B\right)$$

$$\times N(y^{censored}) dy^{miss} dy^{censored}$$

,where $y = \left(y^{obs}, y^{miss}, y^{censored}\right)$, y^{obs} , y^{miss} and $y^{censored}$ represent the observed, missing and censored intensity values respectively; X^{obs} and X^{miss} are the corresponding explanatory variables used in the multivariate multilevel inferential model f for the observed and missing intensities. $N(y^{censored})$ is the normal truncated function for the left-censored values. The parameters B of the multivariate multilevel inference model f, the missing values and the parameters of the missing mechanism are treated as unknowns, and are jointly estimated under a Bayesian framework. The Bayesian approach enables utilization of the prior information learned from the other studies, i.e. Hrydziuszko and Viant [21], as an enrichment of the missing value imputation.

2. A multivariate-multilevel model for a group of proteins without missing intensity

We begin by defining the multivariate framework, initially for the case of complete proteomic data.

In contrast to the traditional single-protein model popularly applied in analysis of proteomics datasets [25], the multiple proteins model analyzes a group of proteins within a multivariate and multilevel framework. Through the use of a two-level model, the multivariate model can be analyzed as a univariate model. In this two-level model, different proteins are treated as level 2 units, and peptide represented by reporter ions are considered the level one units nested within proteins and subjects in the data hierarchy. Ideally the same protein will be observed in all subjects; in such case, proteins and subjects are defined as crossed factors and both are the level 2 units. Although this is in general not the case in practice, the multilevel model allows for unbalances in the design and random missingness [26]. The level one part of the model defines the relationship between the peptides intensities and the experimental factors (MS experiment factors: i.e. Label and nth run). The level two part of the model describes the relationship between the selected protein level random coefficients and the explanatory variables. Another component of the level 2 model defines the relationship between a random intercept for subject and the subject level explanatory variables (clinical design factors: i.e. randomization group and baseline measures).

Define level one part of the 2-level model: A full model with interactions at level one is defined as:

$$
\begin{aligned}
y_{ilp} = {} & \phi_{0l} + \mu_{0p} \times \text{protein}_{ilp} + \mu_{1p}\text{mz}_{ilp} \\
& + \sum_{h=1}^{w} \mu_{2ph}\,\text{label}_{hilp} + \sum_{r=1}^{v} \mu_{3pr}\,\text{run}_{rilp} \\
& + \sum_{h=1}^{w} \mu_{4h}\,\text{mz}_{ilp} \times \text{label}_{hilp} + \sum_{r=1}^{v} \mu_{5r}\,\text{mz}_{ilp} \times \text{run}_{rilp} \\
& + \sum_{k=1}^{wv} \mu_{6k}\,\text{label}_{hilp} \times \text{run}_{rilp} + e_{ilp}
\end{aligned}
\tag{1}
$$

where y_{ilp} denotes the log-transformed intensity for peptide i ($range\ 1{:}n$;), subject l ($range\ 1{:}q$) and protein p ($range\ 1{:}z$). protein_{ilp} denotes the protein identity for p; mz_{ilp} denotes the m/z ratio for peptide i, subject l and protein p; $label_{hilp}$ is a binary variable with value 0 or 1 which indicates response y_{ilp} is from label h ($range1{:}w$); run_{rilp} is a binary variable with value 0 or 1 which indicates response y_{ilp} is from run r ($range1{:}v$) of all experiments; ϕ_{0l} denotes the random intercept

for subject l; $U = (\mu_{0p}, \mu_{1p}, \mu_{2ph}, \mu_{3pr})$ denotes the random coefficients for protein p; $\mu_{4h}, \mu_{5r}, \mu_{6k}$ denote the regression coefficients for the interactions terms of m/z ratio and experimental factors label and run respectively, they are not assumed to vary across proteins; e_{ilp} denotes the unexplained residual error for peptide i, protein p and subject l.

At this level of the model, the intercept can vary across different proteins. The m/z ratio, runs and labels can have different effects on the peptide relative abundance across different proteins, and the intercept can vary across different biological samples.

Define the protein and subject parts of the 2-level model:

    a. The regression coefficients $U = (\mu_{0p}, \mu_{1p}, \mu_{2ph}, \mu_{3pr})$ at the protein level can be described as the second level of the model,

$$\begin{aligned}
\mu_{0p} &= \sum_{c=1}^{g} b_{1pc} \times \boldsymbol{condition}_{lc} + b_{0p} \\
\mu_{1p} &= \mu_1 + b_{2p} \\
\mu_{2ph} &= \mu_{2h} + b_{3ph} \\
\mu_{3pr} &= \mu_{3r} + b_{4pr}
\end{aligned} \qquad , \qquad (2)$$

where $\boldsymbol{condition}_{lc}$ represents a vector of binary variables to identify the physiological condition c ranged between 1 and g for subject l. $b_{0p}, b_{2p}, b_{3ph}, b_{4pr}$ represents the random residuals of protein p for the random intercept, m/z slope, label and run respectively; b_{1pc} represents a random residual of protein p and condition c. $\mu_1, \mu_{2h}, \mu_{3r}$ are the fixed intercepts. The protein part of the level 2 model defines the random regression coefficients of intercept, m/z ratio, label, run and subject's physiological condition. It is assumed that they are different across proteins. There are no fixed terms for the physiological conditions at the protein level as they are separately defined at the subject level in equation (3).

    b. The random intercepts at the subject level can be described separately as

$$\phi_{0l} = \gamma_{00} + \sum_{b=1}^{q} \gamma_{1b} \boldsymbol{Z}_{lb} + \sum_{c=1}^{g} \gamma_{2c} \boldsymbol{condition}_{lc} + c_{0l}, \qquad (3)$$

where q covariates of the subjects are represented by $Z = [Z_{l1}, Z_{l2} \ldots Z_{lq}]$; $\gamma_{00}, \gamma_{1b}, \gamma_{2c}$ represents the fixed intercept and the fixed effect coefficients. c_{0l} represents the residual terms for subject l. The subject part of the level 2 model defines the random intercept for each subject.

Let B be the vector of random effects for proteins, C be the vector of random residuals for subjects, and e be the vector of the random errors. We define the distribution of these random effect vectors as follows:

$$B = (b_{0p}, b_{1pc}, b_{2p}, b_{3ph}, b_{4pr}) \sim MVN(0, \Phi),$$
$$C = (c_{01}, \ldots, c_{0l}, \ldots, c_{0n}) \sim MVN(0, G),$$
$$e \sim MVN(0, R),$$

where Φ denotes the variance-covariance matrix for parameters at the protein level, G denotes the variance-covariance matrix for parameters at the subject level, and R is the variance-covariance matrix for the random residual errors at the peptide level. We assume the random effects for subjects are independent of the random effects for the proteins, and are independent of the random errors at the peptide level. The variance-covariance matrix of responses can be decomposed as $V = K_{n \times z} \Phi_{z \times z} K'_{z \times n} + Z_{n \times q} G_{q \times q} Z'_{q \times n} + R_{n \times n}$. $K_{n \times q}$ is the random effect matrix for protein and $Z_{n \times q}$ is the random effect matrix for subjects, n is the total number of intensity observations in the data set.

3. A mixture model to handle missingness mechanism - a Bayesian approach

After introducing the multivariate model (Equations (1-3)) for complete data, the missingness mechanism is integrated with the same inference model within a Bayesian framework.

3.1. The configuration of the inference model and the imputation model for missingness

In a labelling mass spectrometry experiment, two types of missing peptide data from MS platform are observed; one is identified as zero and the other is identified as blank. The zeros are the left-censored values lower than the detectable limit, and the blanks are missing intensities due to weak signals.

Through the Bayesian framework, we can learn the missing data information from the other studies and incorporate them through the priors of the model. For example, the missing probability of the intensity can be estimated using its associations with the instrumental features of mass spectrometry reported in the other studies. We firstly define the Bayesian multivariate multilevel inference model (4), and secondly define the missing mechanism given by equation (5). Equation (4) describes a two-level

hierarchical model that is a special scenario of equation (1) without interactions. The complete intensity data $\boldsymbol{y} = \left(y^{\text{obs}}, y^{\text{miss}}, y^{censored} \right)$ is assumed to be multivariate normal distributed with mean $\boldsymbol{y}_{i,p}$ and variance-covariance matrix Ω. $y_{ilp}{}^{\text{obs}}, y_{ilp}{}^{\text{miss}}, y_{ilp}{}^{censored}$ represents the observed, missing and censored log-transformed intensity values, of protein p and the subject l respectively.

The inference model:

$$
\begin{aligned}
&\left(\boldsymbol{y}_{ilp} | \text{mz}_{ilp}, \text{run}_{rilp}, \text{label}_{hilp}, \text{condition}_{ilp}, \boldsymbol{B} \right) \\
&\sim MVN(\boldsymbol{y}_{ip}, \Omega) \\
&\boldsymbol{y}_{ip} = \boldsymbol{U}_{0p} + \boldsymbol{U}_{1p} \boldsymbol{mz}^T{}_{ip} + \boldsymbol{U}_{2p} \boldsymbol{condition}^T{}_{ip} \\
&+ \boldsymbol{\beta}_0 \boldsymbol{subject}^T{}_{ip} + \boldsymbol{\beta}_1 \boldsymbol{label}^T{}_i + \boldsymbol{\beta}_2 \boldsymbol{run}^T{}_i
\end{aligned}
\tag{4}
$$

Where $\boldsymbol{B} = (\boldsymbol{U}_{0p}, \boldsymbol{U}_{1p}, \boldsymbol{U}_{2p}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, is the matrix of regression coefficients. $\boldsymbol{U}_{0p}, \boldsymbol{U}_{1p}, \boldsymbol{U}_{2p}$ represent vectors of regression coefficients for different proteins. $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ represent vectors of regression coefficients for different subjects, labels and runs respectively; they are assumed to be the same across different proteins. Ω is a variance-covariance matrix having a common unknown variance term σ^2 across diagonal terms. $\boldsymbol{y}_{ip}$ denotes the mean log-intensity for subject l, $\boldsymbol{y}_{ip}$ is predicted by the covariates of peptides, proteins and subjects in equation (4). $\boldsymbol{mz}^T_{ip}, \boldsymbol{condition}^T_{ip}, \boldsymbol{subject}^T_{ip}, \boldsymbol{label}^T$ and $\boldsymbol{run}^T$ represent the row vectors of explanatory variables m/z, physiological condition, subject, label and run respectively. Prior of the protein level regression coefficients $\left(\boldsymbol{U}_{0p}, \boldsymbol{U}_{1p}, \boldsymbol{U}_{2p} \right)$ is multivariate normal distributed with mean $\boldsymbol{\alpha} = (\alpha_{0p}, a_{1p}, \alpha_{2p})$ and variance-covariance matrix Σ, assuming Σ is the same for all subjects; The hyper-prior of $\boldsymbol{\alpha}$ is multivariate normal distributed with mean $\boldsymbol{f}$ and variance-covariance $\boldsymbol{G}$.

A well-known hyper-prior for Σ is inverse-Wishart distribution [27], $\Sigma^{-1} \sim \text{Wishart}(\boldsymbol{R}, d = 3, v_0)$, where $\boldsymbol{R}$ is a positive scaled definite matrix and $v_0 > d - 1$. We can assign non-informative or informative hyper-priors for vector $\boldsymbol{f}$, matrices $\boldsymbol{G}$ and $\boldsymbol{R}$. Priors for the residual variance σ^2 and the random intercept for subject $\boldsymbol{\beta}_0$ and regression coefficients $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ are given as follows:

$$
\sigma^2 \sim IG(a, b), \boldsymbol{\beta}_1 \sim \text{N}(0, d_1), \boldsymbol{\beta}_2 \sim \text{N}(0, d_2), \boldsymbol{\beta}_0 \sim N(0, d_3),
$$

where σ^2 is inverse-Gamma distributed with shape a and scale b, $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ are normal distributed, non-informative priors can be assigned to $a, b, d_1, \dots d_3$ if no prior information is available.

3.2 The missing mechanism

Let r be a binary indicator variable denoting if the intensity has a missing value (1: missing, 0: not-missing). Based on observations [Figure 1] and prior information, we know that the probability of missingness associated with the intensity value and the m/z ratio. If we assume that the probability of intensity being missed (including censored) is $pm = \{pm_{ilp}\}$ (for $r_{ilp} = 1$), and it is Bernoulli distributed. The missing data mechanism is defined in the following equation (5):

$$\text{logit}\left(pm_{ilp}\right) = \theta_0 + \theta_1 \times mz_{ilp} + \theta_2 \times y_{ilp} \tag{5}$$

Let $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)$ be the vector of missing data mechanism parameters. The conditional posterior distribution of $\theta_0, \theta_1, \theta_2$ is postulated as follows:

$$g_4(\theta_0, \theta_1, \theta_2 | \boldsymbol{mz}, \boldsymbol{y}, \boldsymbol{r}) = \prod_{j=0}^{2} g_{4j}\left(\theta_j | \boldsymbol{mz}, \boldsymbol{y}, \boldsymbol{r}\right)$$

$$\propto f_2(\boldsymbol{r} | \boldsymbol{mz}, \boldsymbol{y}, \theta_0, \theta_1, \theta_2) \times \prod_{j=0}^{2} g_{4j}\left(\theta_j\right),$$

where $\boldsymbol{mz} = \{mz_{ilp}\}$ represents the vector of m/z ratios, g_{4j} represents the density function for the priors of θ_j ($j = 0,1,2$). $f_2(\boldsymbol{r} | \boldsymbol{mz}, \boldsymbol{y})$ is the logistic regression likelihood function of equation (5). Because the missing peptide intensity is not observable in the data set, and no information can be obtained from the observations, a pair of informative priors is used based on the monotone relationship expected between missingness and intensity.

The censored values $y_{ilp}^{censored}$ are proposed to belong to a truncated normal distribution with an upper threshold of c, $y_{ilp}^{censored} \sim N(\boldsymbol{y}^{censored} | c)$.

3.3 Hamiltonian Monte Carlo (HMC) and Non-U-Turn posterior sampling (NUTs) algorithm

Our proposed method utilizes Hamiltonian MC/Non U Turn Sampling [28] for the posterior distribution. Compared to the Gibb sampling, Hamiltonian MC (HMC) [29] avoids the random walk by introducing the leapfrog function. It provides an alternative to approximating the solution on the continuous time scale from the solutions on the discrete time scale, with a specified step size. The logarithmic posterior probability function was simulated by one of the paired partial differentiated equations, namely the Hamiltonian. Larger moving steps are generated from the leap frog scheme, and

this helps to improve the convergence compared to the random walk. It has been shown to have higher efficiency in sampling high-dimensional correlated multivariate distributions.

RStan is a new tool recently developed to implement HMC modelling for Bayesian data analysis, of which the posterior distribution can be sampled using the No-U-Turn Sampling method (NUTS)-an extension of HMC. NUTS [28] implements a recursive algorithm that will enable auto-tuning of the coupling parameters in HMC: numbers of leap-frog steps and the discrete step. In RStan, a known data variable cannot have missing values. If a variable contains missing values, the missing values will need to be defined separately as an unknown variable. The RStan algorithm and program for the model described in equation (4-5) are given in 3.4.

3.4 The program of the RStan HMC/NUTS for the missing data model (with only label as the covariate in this demonstration)

The RStan program for the proposed model comprises five blocks: data, transformed data, parameters, transformed parameters and model.

data

Step 1: Assign variables names to all the observed peptide intensities and their corresponding explanatory variables in the data block, excluding records with censored and missing intensities;

Step 2: Assign different variables names to all the observed explanatory variables linked to the censored and missing intensities in the data block;

Step 3: Assign priors and hyper priors of the covariance matrix for protein level covariates, and other known values (i.e. number of subjects, number of explanatory variables) in the data block.

```
data
    {
    int nobs;                              //number of records
    int ncensor;                           //number of censored
    int nprotein;                          //number of protein
    int nmiss;                //number of missing
    int  nsubject;                                    //number  of  subject

    int<lower=0> subject[nobs];            //subject id
    int<lower=0> proteinid[nobs];      //protein id
    int<lower=0,upper=1> classno[nobs];    //subject group
    int<lower=0,upper=1> miss[nobs];          //missing value indicator
    …
    corr_matrix[3] R;              //hyper prior for scaled covariance matrix
    vector[3] Sigma_sd;            //hyper prior for diagonal term of covariance
```

```
matrix
      cov_matrix[3] prec;        //hyper prior for precision matrix of latent covariance
matrix
      vector[3]  mn;                   //hyper  prior  for  mean  of  the  latent  protein
parameters
      }
```

transformed data

Any data needing to be transformed in the program is given in the transformed data

block. For instance, to derive a prior for the covariance matrix if the prior is given as a

precision matrix in the data block.

```
transformed data
      {
      cov_matrix[3] T;                //prior covariance matrix for level 2 parameters
      cov_matrix[3] invprec;          //prior covariance matrix for latent covariance
matrix
      T<-diag_matrix(Sigma_sd) * R * diag_matrix(Sigma_sd);
      invprec<-inverse(prec);
      }
```

parameters

Any unknown variables including the missing and censored intensities, and variables

that will be used to re-parameterize the unknown variables, are defined in the

parameter block. The re-parameterization is to improve the statistical efficiency in

posterior samplings [30].

Variables used for re-parameterizations are named as latent variables. In the NUTS

program in the supplementary material, the latent parameters defined in this block are

named by adding "latent" to the unknown parameters. For example, beta4_ latent is

the latent variable for regression coefficient beta4.

```
parameters
      vector[3] U_latent[nprotein];
      vector[3] g;
      cov_matrix[3] pVAR;
      real<lower=0> ita;

      real alpha_latent;
      real alpha_mu;
      real<lower=0> alpha_theta;

      real alpha1_latent;
      real alpha2_latent;
      vector[64] beta2_latent;
      real<lower=0> beta2_theta;
      vector[64] beta2_mu;

      vector[4] beta4_latent;
      real<lower=0> beta4_theta;
```

```
        vector[4] beta4_mu;

        vector[nmiss] logofAUC_m_latent;
```

transformed parameters

Any unknown variable that comprises both known and unknown values can be derived in the transformed parameters block. The missing and censored peptide quantities and the probability of missing are derived in this block. The parameters use re-parameterization techniques are also derived from this block.

```
transformed parameters
for (prot in 1:nprotein)
        U[prot]<-g+pVAR*U_latent[prot];                    //derived   re-parameterized
parameter

        alpha<-alpha_mu+alpha_theta*alpha_latent;   //   derived   re-parameterized
parameter
        alpha1<-0.0085+2.5e-7*alpha1_latent;          //  derived   re-parameterized
parameter
        alpha2<--0.45+0.4*alpha2_latent;              //  derived   re-parameterized
parameter

//derivation for intensity and probability of missing
        for (pep in 1:nobs)
                {
mu[pep]<-beta2[subject[pep]]+U[protein_id[pep]][1]+U[protein_id[pep]][2]*mass_to_c
harge_ratio
[pep]+U[proteinid[pep]][3]*classno[pep]+beta4[1]*label1[pep]+beta4[2]*label2
[pep]+beta4[3]* label3[pep]+beta4[4]* label4[pep];
  pmiss[pep]<-inv_logit(alpha+alpha1*mass_to_charge_ratio
[pep]+alpha2*intensity[pep]);
                }
```

model

In the model block, all the unknown parameters are defined by its sampling distribution using ~.

Step 1: Sampling for the define priors and hyper priors: distributions of all priors or hyper priors for unknown parameters are defined firstly in the model block.

Step 2: Sampling for the latent variables used for re-parameterization: the re-parameterization is equivalent to a two-steps sampling that firstly samples the latent parameters (i.e., location and scale) according to their hypothesized distributions in the model block; and secondly derives the unknown parameters using the selected latent parameters in the transformed parameters block.

Step 3: Missing and censored data parameters: The missing and censored values are

treated as a couple of unknown variables. Logistic regression is used to model the missing mechanism and is defined in the transformed parameter block. The posterior distribution of missing values is sampled through the re-parameterization. The posterior distribution of the censored value is estimated by integration of a truncated normal distribution.

```
Model

        beta2_latent~normal(0,1);    //sampling for latent variable for regression
coefficient
        beta2_mu~normal(0,1);        //sampling for mean of the latent variable
        beta2_theta~gamma(1,1);    //sampling for variance of the latent variable

        beta4_latent~normal(0,1);   //sampling for latent variable for regression
coefficient
        beta4_mu~normal(0,1);    //sampling for mean of the latent variable
        beta4_theta~gamma(1,1); //sampling for variance of the latent variable

        g~multi_normal(mn,T);          //sampling for latent variable for protein
parameters

        pVAR~inv_wishart(3,invprec); //sampling for latent covariance matrix


        for (prot in 1:nprotein)
        U_latent[prot]~multi_normal(mn,R);       //standard multinormal distributed
        ita~gamma(1,1);                              //prior for variance of error term
        logofAUC~normal(mu,ita);                 //intensity distribution

        //sampling for latent parameters in missingness model
        alpha_latent~normal(0,1);
        alpha_mu~normal(0,1);
        alpha_theta~gamma(1,1);
        alpha1_latent~normal(0,1);
        alpha2_latent~normal(0,1);
        miss~bernoulli(pmiss);
        miss_m~bernoulli(pmiss_m);
        logofAUC_m_latent~normal(0,1);

    //censored left truncated normal distribution
    for (pep in 1:ncensor)
    target +=log(Phi((censor_lim-mu_cen[pep])/ita)+0.001);
```

4. Using simulation to validate the model for the missing mechanism and compare between *lmer,* BUG and NUTs estimation

An iTRAQ proteomic experiment with complete data was simulated; the same dataset with non random missing patterns as described in [21] was added in the same dataset. The estimates from three different methods (lmer, BUG and NUTs) were compared against the known parameters from the complete data.

The simulated study has a row-column design for eight runs (row), eight labels

(column) and two classes that comprised 32 subjects from the healthy population as controls and 32 subjects from the diseased population. A Poisson distribution ($\lambda=5$) was used to generate the number of peptides of 200 proteins. The intensity of the peptide has confounding effects of run, label, the total amount of protein in fluid tissue sample, subject class, protein intercept and mass-to-charge ratio (m/z). The non-random missingness pattern are associated with the peptide intensity and the mass-to-charge ratio (m/z). The censored threshold was set to 0.1 which was the minimal log-intensity of an iTRAQ case study. Any simulated log-intensity below this value was set to zero. The simulated data set of 78400 records has 928 (1.2%) observations with censored peptide intensities and 14287 (18.2%) observations with completely missing peptide intensities.

We compared the univariate ANOVA, a multivariate mixed model fitted with *lmer* in R [31] using complete data only; we also compared the multivariate *lmer* estimates with the MNAR Bayesian model estimates that including observed covariates of missing intensities.

Using complete data, estimates of mean class differences (diseased vs. control) from the multiple proteins R/*lmer* model demonstrated a better agreement with the actual values, than estimates from the univariate single protein models did (scatter plots figure 2). In two sets of similar linear regressions that include actual values as dependents, estimates of single or multiple protein models are included as independents respectively comparing against the actual values. The R^2 is 0.87 with a slope of 1.21 (standard error: 0.034) in the regression using estimates from the univariate single protein model. The R^2 is 0.91 with a slope of 1.27 (standard error: 0.028) in the regression using estimates from the multiple proteins mode. The R^2 is larger and standard error of the slope is slightly smaller using estimates from the multiple proteins model.

When the non-random missingness mechanism is modelled using the Bayesian method, a program using BUGS and a program using HMC/NUTS were used to derive the posterior distributions of unknown parameters. Both programs have smaller variability in the estimates of the mean class difference in comparison to the actual values (scatter plot figure 3). The HMC/NUTS program, which models the censored value as normal truncated distributed, achieves a better agreement with the actual values. This is indicated by the closer-to-1 slope of the linear regression that

has the HMC/NUTS estimates compared against the actual values. The achieved R^2 is 0.95, and the slope is 1.07 (standard error: 0.018).

The unexplained residual variance of the model using the HMC/NUTS algorithm is 1.87 (95% credible interval: 1.81, 1.93), which is similar to the actual value of 2.0, and smaller than 3.11- the result of the R *lmer* model of the complete cases [Figure 3].

5. A cardiac proteomic case study

The cardiac proteomic study is one part of a double blinded randomized control trial [32] that investigated how the intra-coronary metoprolol (beta blocker) changed the myocardial metabolism of peptides, proteins and metabolites profiles in patients admitted to hospital with a first myocardial infarction. The cardiac proteomic study investigated how the whole plasma proteome profile changed after coronary angioplasty intervention. The study collected five plasma samples at different time points in eight patients during their percutaneous coronary interventions. Two of these plasma samples (before and 20 minutes after a controlled coronary occlusion) were selected in the proteomic study. The IPI human database (European Bioinformatics Institute, Hinxton, UK) was used in the Protein Pilot Software to match the observed peptides with their corresponding proteins for the identification.

This case study has 17780 peptide observations of which 1667 (9.4%) had censored and 218 (1.2%) had completely missing values in the peptide intensities. The Bayesian model with the selection model for the missing mechanism is constructed as follows:

$$y_{ilp} = \phi_{0l} \, subject_{ilp} \quad + \beta_{0p} protein_{ilp} + \beta_{1p} protein_{ilp} \times mz_{ilp}$$

$$+\beta_{4p} protein_{ilp} \times sampling\_time_{ilp} + \sum_{h=1}^{4} \beta_{2h} \, label_{hilp}$$

$$+\sum_{r=1}^{8} \beta_{3r} \, run_{rilp}$$

$$\text{logit} \, (pm = \text{Pr} \, (r)) = \alpha_0 + \alpha_1 mz_{ilp} + \alpha_2 y_{ilp}$$

where y_{ilp} defines the mean for the peptide intensities including complete, completely missing and censored values on the log scale, $y_{ilp} \sim N(\mu_p, \sigma^2)$. $subject_{ilp}$ defines the subject identity and ϕ_{0l} defines the intercepts for each subject. $\beta_{0p}, \beta_{1p}, \beta_{4p}$ define the protein level parameters: intercept, m/z ratio and

sampling time. The difference in log-intensities between sampling time is equivalent to the intervention effect. $protein_{ilp}$ is a variable recording the identity of protein p. β_{2h} and β_{3r} are regression coefficients for label h and run r respectively. pm defines the probability of having a missing intensity value; and $\alpha_0, \ldots, \alpha_1$ define the regression coefficients in the logistic regression for pm.

The coefficients of logistic regression α_1 and α_2 are part of the joint unknown parameters with a pair of chosen normal distributed informative priors $N(0.0085, 4 \times 10^{-8}), N(-0.45, 0.25)$ for m/z ratio and peptide abundance respectively [21]. The protein level parameters $\beta_{0p}, \beta_{1p}, \beta_{4p}$ used a multivariate normal distributed prior (**y**, *T*) where,

$$\boldsymbol{y} \sim \boldsymbol{MVN}(\boldsymbol{\mu}, \Omega^{-1}), \mathrm{T} \sim \mathrm{invWISHART}(\Phi^{-1}, \mathbf{3})$$
$$\boldsymbol{\mu} = (\mathbf{0.5}, \mathbf{0.5}, \mathbf{0.5}),$$
$$\Omega = \begin{pmatrix} \mathbf{0.01} & \mathbf{0.01} & \mathbf{0.01} \\ \mathbf{0.01} & \mathbf{0.01} & \mathbf{0.01} \\ \mathbf{0.01} & \mathbf{0.01} & \mathbf{0.01} \end{pmatrix},$$
$$\Phi = \begin{pmatrix} \mathbf{0.01} & \mathbf{0.001} & \mathbf{0.001} \\ \mathbf{0.001} & \mathbf{0.01} & \mathbf{0.001} \\ \mathbf{0.001} & \mathbf{0.001} & \mathbf{0.01} \end{pmatrix}$$

The non-informative priors for the peptide level parameters $\beta_{2,h}$ and $\beta_{3,r}$ are normal distributed and denoted as $N(0,10)$ with mean 0 and precision 0.1, and the non-informative prior for the subject level parameters $\phi_{0,l}$ is normal distributed and denoted as $N(0,1)$ with mean 0 and precision 1.

Using the complete data, the significant fixed effect of the intervention derived from the R/*lmer* model reveals a systematic shift of -1.42 in all protein expressions on the log-scale; the other significant fixed effects include the centralized m/z, label 116 and 117 (table1.). The predicted random effects of intervention for different proteins indicate the magnitudes of fold changes in the protein abundances introduced by the intervention (figure 5). Compared to the variance components in R/lmer model, using NUTS model results in a similar intervention variance but a larger m/z ratio variance. The unexplained residuals variance in the NUTS and BUGS models are 1.62 and 0.37 respectively (table 2); both are smaller than 2.57 in the R/*lmer* multiple protein models.

6. Discussion

The multilevel model proposed in this paper provides a new approach to analysing a group or all groups of proteins when protein abundances comprise missing values that are Missing-Not-At-Random. The flexibility of multivariate multilevel model allows for an unbalanced design structure among the responses. This becomes an advantage in a proteomic study when the numbers of component peptides of the same protein are unequal across different subjects. Within this multivariate framework, experimental factors of multiple proteins can be utilized in conjunction with clinical design factors to derive the results of a single protein. Inclusion of covariates from multiple proteins improves precision of the parameters at protein level and results in shrunk estimates for a single protein with few observations. The multivariate model also enables identification of proteome-wide characteristics, such as systematic differences in protein abundances between two sampling times in the cardiac case study.

When non-random missingness needs to be considered and modelled, using numerical integration or an EM algorithm is not feasible in such a complex framework. Using Hamiltonian MC/NUTS is an improvement, especially in the case of utilizing non-standard distribution for handling missing data. Compared to BUGS, RStan is more flexible for handling missing data via the truncated distribution function. The NUTS algorithm is more efficient in dealing with the truncated distribution and non-standard distribution than BUGS.

The philosophy of the proposed multilevel multivariate method can be generalized to other proteomics or "omic" research with different experimental structures. Future development of the model will include consideration of various covariance matrices for protein and subject level variables, and adding known factors as explanatory variables to estimate components of the covariance matrices.

Acknowledgements

References

1. Silva, J.C., et al., *Quantitative proteomic analysis by accurate mass retention time pairs.* Analytical Chemistry, 2005. **77**(7): p. 2187-200.
2. Stephens, A.N., P. Quach, and E.J. Harry, A streamlined approach to high-throughput proteomics. *Expert Review of Proteomics*, 2005. **2**(2): p. 173-85.
3. Vitzthum, F., et al., Proteomics: from basic research to diagnostic application. A review of requirements & needs. *Journal of Proteome Research*, 2005. **4**(4): p. 1086-97.
4. Rodriguez , H., et al., Recommendations from the 2008 International Summit on Proteomics Data Release and Sharing Policy: the Amsterdam principals. *Journal of Proteome Research*, 2009. **8**(7): p. 3689-3692.
5. Maddalo, G., et al., Analytical assessment of MALDI-TOF Imaging Mass Spectrometry on thin histological samples. An insight in proteome investigation. *Clinica Chimica Acta*, 2005. **357**(2): p. 210-8.
6. Pelzing, M. and C. Neususs, Separation techniques hyphenated to electrospray-tandem mass spectrometry in proteomics: capillary electrophoresis versus nanoliquid chromatography. *Electrophoresis*, 2005. **26**(14): p. 2717-28.
7. Shen, Y. and R.D. Smith, Advanced nanoscale separations and mass spectrometry for sensitive high-throughput proteomics. *Expert Review of Proteomics*, 2005. **2**(3): p. 431-47.
8. Palmblad, M., A. Tiss, and R. Crame, Mass spectrometry in clinical proteomics - From the present to the future   *Proteomics - Clinical Applications*, 2009. **3(1)**: p. 6-17
9. Kline, K.G. and M.R. Sussman, Protein quantitation using isotope-assisted mass spectrometry. *Annual Review of Biophysics*, 2010. **39**: p. 291-308.
10. Corthals, G.L. and K. Rose, *Quantitation in Proteomics*, in *Proteome Research: Concepts, Technology and Application*. 2007, Springer: Berlin.
11. Kiyonami, R., T. Schlabach, and K. Miller, *Identification and quantification of iTRAQ labeled peptides on the Finnigam LTQ using MS/MS and MS*. 2005, Thermo electron coporation: application notes
12. Shadforth, I.P., et al., i-Tracker: for quantitative proteomics using iTRAQ. *BMC Genomics*, 2005. **6**: p. 145.
13. Chong, P.K., et al., Isobaric tags for relative and absolute quantitation (iTRAQ) reproducibility: Implication of multiple injections   *Journal of Proteome Research*, 2006. **5(5)**: p. 1232-1240.
14. D'Ascenzo, M., l. Choe, and K.H. Lee, iTRAQPak: an R based analysis and visualization package for 8-plex isobaric protein expression data. *Briefings in Functional Genomics and Proteomics*, 2008. **7(2)**(No 2.): p. 127-135.
15. Hamdan, M. and P.G. Righetti, Modern strategies for protein quantification in proteome analysis: advantages and limitations. *Mass Spectrom Review*, 2002. **21**(4): p. 287-302.
16. Sechi, S. and Y. Oda, Quantitative proteomics using mass spectrometry. *Current Opinion in Chemical Biology*, 2003. **7**(1): p. 70-77.
17. Boehm, A.M., et al., Precise protein quantification based on peptide quantification using iTRAQ™. *BMC Bioinformatics*, 2007(8).
18. Wiese, S., et al., *Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. Proteomics*, 2007(7): p. 340-350.
19. Leitner, A. and W. Lindner, Current chemical tagging strategies for proteome analysis by mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci*, 2004. **25**(813): p. 1-26.
20. Breitwieser, F.P., et al., General Statistical Modeling of Data from Protein Relative Expression Isobaric Tags. *Journal of Proteome Research*, 2011. **10**: p. 2758-2766.
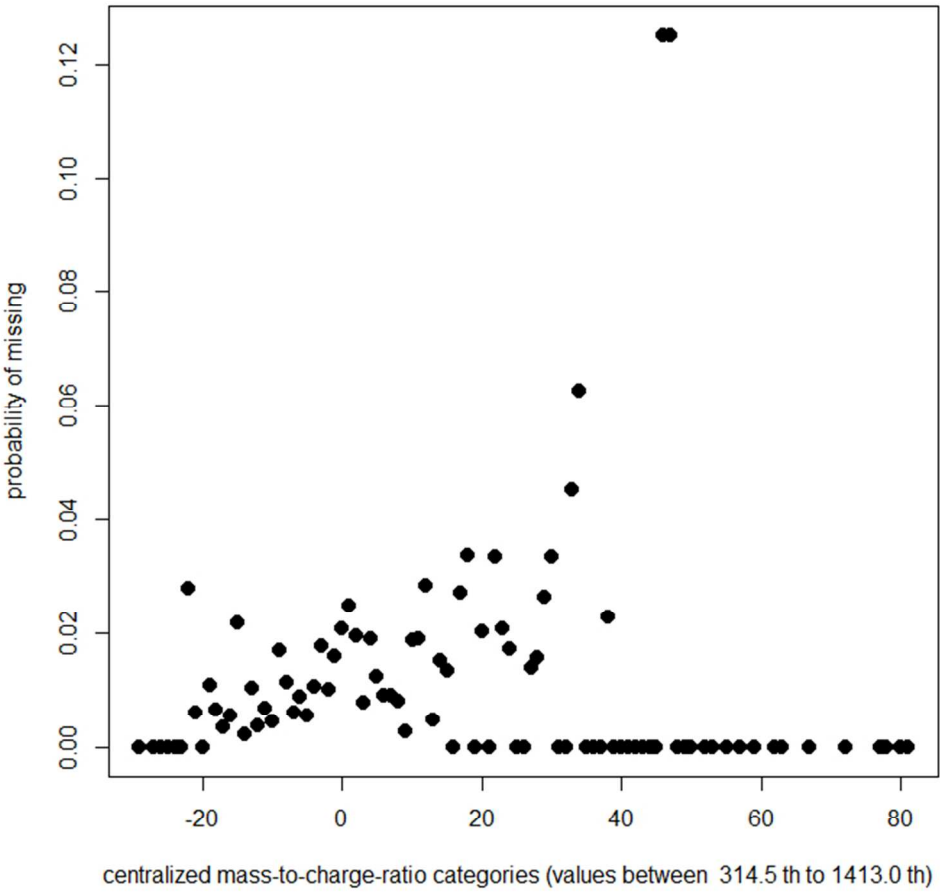
21. Hrydziuszko, O. and M.R. Viant, Missing values in mass spectrometry based metabolomics:an undervalued step in the data processing pipeline. *Metabolomics*, 2012. **8**: p. S161–S174.

22. Wells, G., H. Prest, and Russ IV, I. Agilent Technologies, Editor. 2011, Agilent Technologies, Inc. : U.S.A.

23. Little, R.J.A., Pattern-mixture models for multivariate incomplete data. *J. Am. Statist. Assoc.*, 1993. **88**: p. 125-134.

24. Little, R.J.A. and D.B. Rubin, *Statistical Analysis with Missing Data.* 1987, New York John Wiley.

25. Hill, E.G., et al., A Statistical Model for iTRAQ Data Analysis. *Journal of Proteome Research*, 2008. **7**: p. 3091-3101.

26. Goldstain, H., *Multilevel Statistical Models*. 1995, London;E. Arnold; New York: Halsted Press

27. Gelman, A. and J. Hill, Data analysis using regression and multilevel/hierarchical models. *Cambridge University Press*, 2007.

28. Hoffman, M.D. and A. Gelman, The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of machine learning research*, 2011. **12**.

29. Neal, R.M., *MCMC using Hamiltonian dynamics*, in *Handbook of Markov Chain Monte Carlo*, S. Brooks, et al., Editors. 2011, Chapman & Hall/CRC: U.S.A.

30. Team, S.D., *Stan Modeling Language: User's Guide and Reference Manual. Version 2.12.0.* 2016.

31. Pinheiro, J., et al., *nlme: Linear and Nonlinear Mixed Effects Models*. 2014: http://CRAN.R-project.org/package=nlme.

32. Erogbogbo, F., et al., Bioengineering Silicon Quantum Dot Theranostics using a Network Analysis of Metabolomic and Proteomic Data in Cardiac Ischemia. *Theranostics*, 2013. **3**(9): p. 719-728.

Table 1. The multivariate multilevel model of R/lmer for the cardiac case using complete data

Random effects:

Groups	Name	Variance	Std. Dev.	Corr	
protein	(Intercept)	0.09	0.29		
	centralized m/z	7×10^{-7}	0.0009	-0.053	
	intervention	0.52	0.72	-0.64	0.205
subject	(Intercept)	1.50	1.23		
Residual		2.57	1.60		

Number of observations: 15895, groups: protein, 151; subject, 8

Fixed effects:	Estimate	Std. Error	t value
(Intercept)	3.90	0.87	4.50
run2	-0.012	1.22	-0.010
run3	-0.0087	1.22	-0.007
run4	-1.50	1.23	-1.23
label115	-0.07	0.04	-1.66
label116	0.43	0.05	8.22
label117	-0.22	0.047	-4.68
centralized m/z	-0.0015	0.0001	-10.15
intervention	-1.42	0.074	-19.25

Table 2. The MNAR Bayesian model parameters

Gibbs:			**NUTS:**		
Posterior of the protein level variance components			Posterior of the protein level variance components		
Groups	**Name**	**Variance (median(IQR))**	**Groups**	**Name**	**Variance (median(IQR))**
Protein	(Intercept)	0.053 (0.081,0.034)	Protein	(Intercept)	0.32 (0.29, 0.34)
	centralized m/z	8.3e-5 (6.6e-5,1.1e-4)		centralized m/z	0.036 (0.031,0.037)
	intervention	0.032 (0.021,0.049)		intervention	0.73 (0.65, 0.75)
Residual		0.37 (0.36, 0.38)	Residual		1.61 (1.608,1.62)
Posterior distribution of the logistic regression coefficients for the missingness			Posterior distribution of the logistic regression coefficients for the missingness		
Groups	**Name**	**Coefficient (median (2.5%-97.5%))**	**Groups**	**Name**	**Coefficient (median(IQR))**
Intercept		-8.20 (-10.14,-7.25)	Intercept		-4.24 (-4.29,-4.18)
m/z		0.0084 (0.0084, 0.0088)	m/z		0.0085 (0.0085, 0.0085)
peptide abundance		-2.89 (-2.29,-1.97)	peptide abundance		-0.76 (-0.76,-0.70)

centralized mass-to-charge-ratio categories (values between 314.5 th to 1413.0 th)

Probability of having missing value verse range of centralized mass-to-charge ratio

244x250mm (72 x 72 DPI)
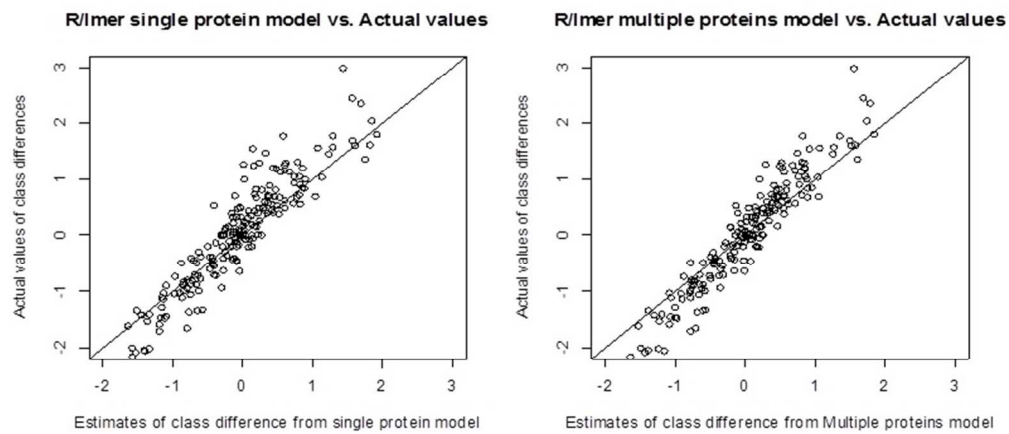
The comparisons in the estimates of the class differences from R/lmer. Left: single protein model vs. the actual values. Right: multiple protein model vs. the actual values.
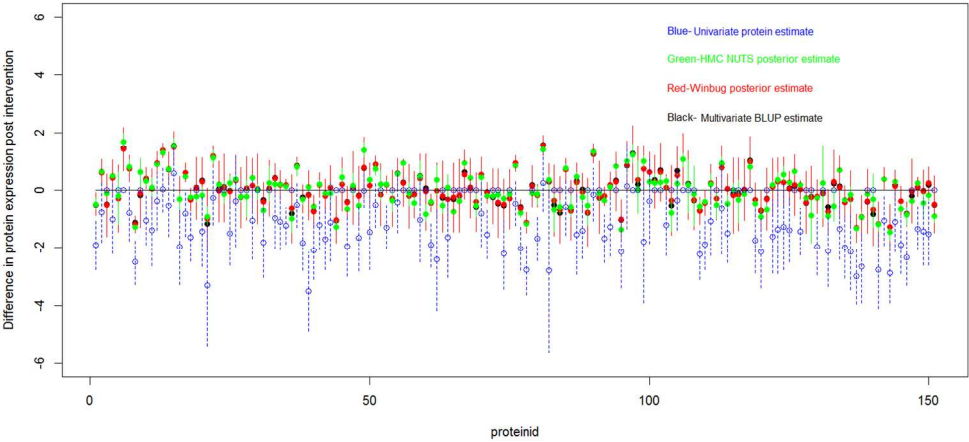
168x79mm (120 x 120 DPI)
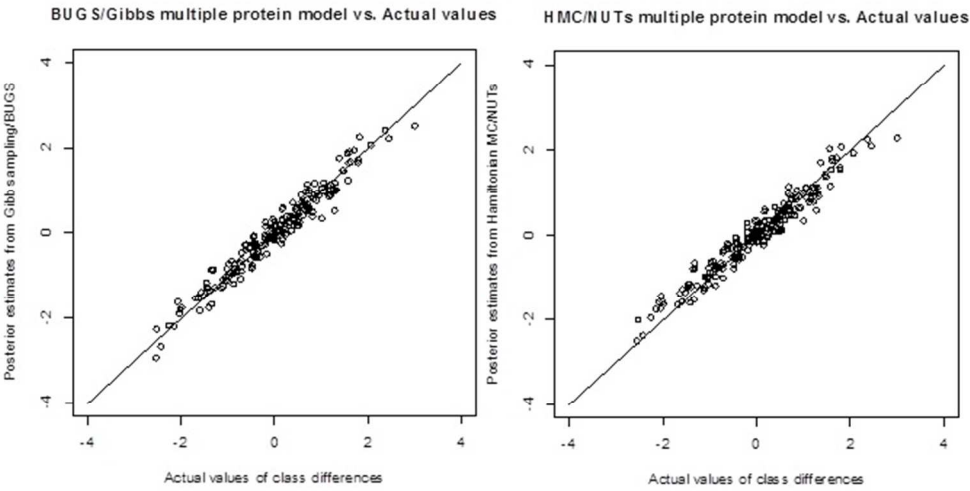
The comparisons in the estimates of the class differences from Bayesian inferences. Left: BUGS/Gibbs model vs. the actual values. Right: HMC/NUTs model vs. the actual values.

145x77mm (120 x 120 DPI)
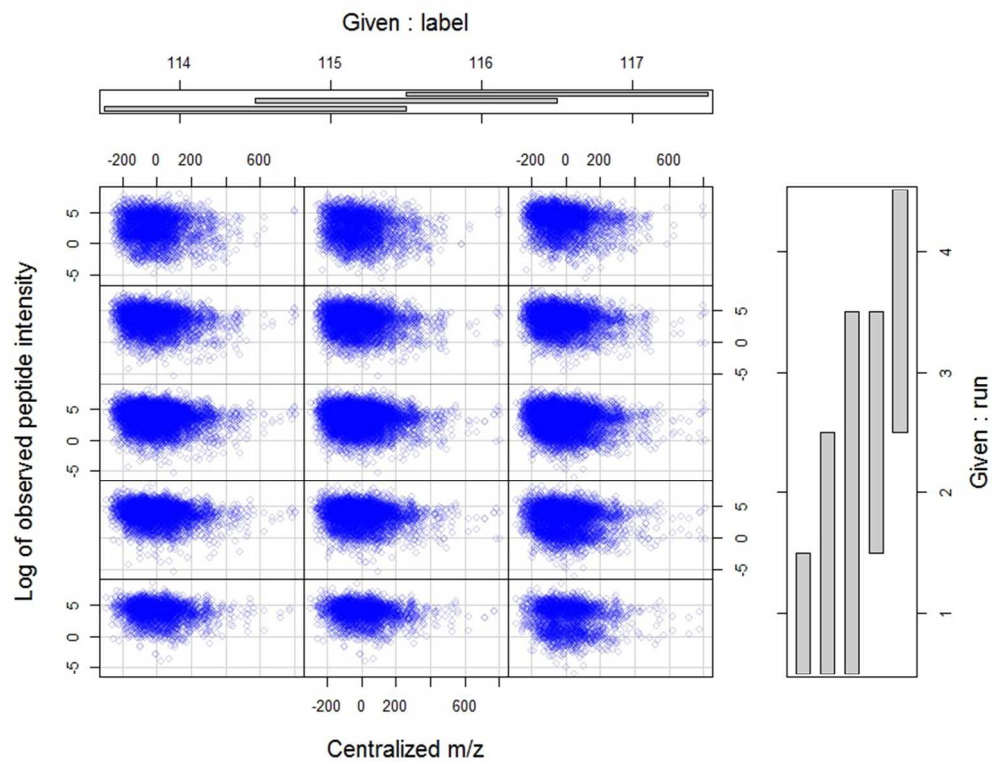
The conditional plots of centralized m/z and peptide intensity given run and label.

182x140mm (120 x 120 DPI)

Error bar plots of the fold change for all proteins

411x212mm (96 x 96 DPI)