Estimation and Inference in Metabolomics with Nonignorable Missing Data

by

Shangshu Zhao

Bachelor of Science, ShanghaiTech University, 2019

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This thesis was presented

by

Shangshu Zhao

It was defended on

April 2nd 2021

and approved by

Christopher McKennan, Department of Statistics

Satish Iyengar, Department of Statistics

Tingting Zhang, Department of Statistics

Copyright © by Shangshu Zhao
2021

Estimation and Inference in Metabolomics with Nonignorable Missing Data

Shangshu Zhao, M.S.

University of Pittsburgh, 2021

Mass-Spectrometry(MS) is one of the most important methods used to characterize metabolomics data. However large-scale MS metabolomics data always faces the problem of data points unobserved or lost, whose magnitude could reach a level where it can't be simply ignored. To account for the information hidden within missing values, we developed a methods to analyze metabolomics data with missing data based on MetabMiss, a newly developed rigorous method to model missing values. Our methodology shows an overall better performance on estimating both coefficients and variance and other criteria, which gives us advantages doing further statistical inference. For each criteria, we demonstrate our method on a simulation data set to compare it to other classical methods.

Keywords metabolomics, missing not at random (MNAR), PCA, WGCNA.

Table of Contents

List of Figures

Preface

This thesis is under the supervision of Professor Christopher Mckennan. I want to specially thank him for the great amount of patience and passion he shows all the time since we've met. The things I have learned from him is not only the lectures, but also the spirit to encourage myself to stay hungry, stay foolish. I will always remember this valuable course from him.

I dedicate my dissertation work to my family and many friends. A special feeling of gratitude to my loving parents, Xiuzhen Shang and Chuanliang Zhao whose words of encouragement and push for tenacity ring in my ears. I would like to express my deep gratefulness to them for all the things they've given up for me. I wouldn't dream to fully understand you no matter what.

Finally, I dedicate this work and give special thanks to my girlfriend Liying Chen for being there for me whenever I need you. Your accompany and support are something I was not even dream for. You showed up in my life in a way that I believe this is what destiny should look like. It honesty feels like it has been years with you. I look forward to be in the same city with you, everyday.

1.0 Introduction

With the development of biology characteristic technology, metabolomics data are becoming more and more easy to access, which gives researchers more opportunities to analyze bio-activities on metabolites level. However, large-scale metabolomics data always faces several problem that makes it not easy to be directly analyzed.

The main problem is that there will always exist a large amount of missing values in the data set, of which a considerable portion are missing at 50% or even more. Thus, simply ignoring the missing values will significantly reduce the total information we are able to utilize from the data set. On the other hand, missing value themselves could be considered as a kind of information that maybe related to underlying true data. From this intuition combined with our understanding to experimental equipment, imputation is developed to fill in the missing value by varies ways, from which we are able to analyze the data. However, imputation methods highly depend on the structure of true metabolomics data, which leaves a paradox that we cannot decide the data structure before analysis.

To get a more robust solution for this problem, MetabMiss [4] propose a novel method that basically assumes missing probabilities are from a certain distribution which is related to the underlying true value of metabolomics concentration, as well as other control parameters. This gives them a way more robust and efficient way to evaluate the missing probabilities. Based on their work, we took a closer look at metabolomics data, and found that there are more structure information that we are able to harvest from. We then developed a novel method that can significantly improve our estimation to the coefficients of interest compared to MetabMiss.

2.0 Problem Setup

Estimating coefficients of interest has always been difficult since the existence of missing value. Thanks to MetabMiss from [4], we are able to get an efficient and useful modeling method. Based on MetabMiss and a bit more assumptions on the In this chapter, I'll introduce how we model the MS metabolomics data with missing values.

2.1 Notation

We use $\mathbf{1}_n, \mathbf{0}_n \in \mathbb{R}^n$ to denote all ones and all zeros vector, $I_n \in \mathbb{R}^n$ to denote identity matrix, with $n > 0$ a integer, $[n]$ the set $\{1, \ldots, n\}$. In terms of the vector and matrix, $\boldsymbol{x}_i$ denotes the i^{th} element of vector $\boldsymbol{x} \in \mathbb{R}^n$, and $\boldsymbol{M}_{ij}$ denote the $(i,j)^{th}$ element of matrix $\boldsymbol{M} \in \mathbb{R}^{n \times m}$. $\mathcal{N}(\cdot; \mu, \tau)$ is the likelihood function for normal distribution with mean μ and variance τ.

2.2 A Description of Data Generating Model

We let y_{gi} denote the observed or unobserved log-transformed metabolite integrated intensity for metabolite $g \in [p]$ in sample $i \in [n]$, which is proportional to a metabolite's concentration in sample. Let $\boldsymbol{X} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n \times d}$ and $\boldsymbol{C} = (c_1, \ldots, c_n)^\top \in \mathbb{R}^{n \times K}$ be observed and unobserved covariates (i.e. latent factors), where the former may contain biological factors like personal disease history, and technical factors that may be related to experiment condition. Therefore the mean model for y_{gi} is given by:

$$y_{gi} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_g + \boldsymbol{c}_i^\top \boldsymbol{\ell}_g + \boldsymbol{\epsilon}_{gi}, \quad \boldsymbol{\epsilon}_{gi} \sim \mathcal{N}(0, \sigma_g^2), \quad g \in [p]; i \in [n]. \tag{1}$$

For the residuals term, we assume that the $\{e_{gi}\}_{g \in [p], i \in [n]}$ are independent and $\{e_{gi}\}_{i \in [n]}$ are identically distributed for each $g \in [p]$. Compared to the model in [4], which assumed a more

general model for the residuals, here we assume the residuals to be normally distribution. This assumption could be justified by general metabolomics data as we plot below:

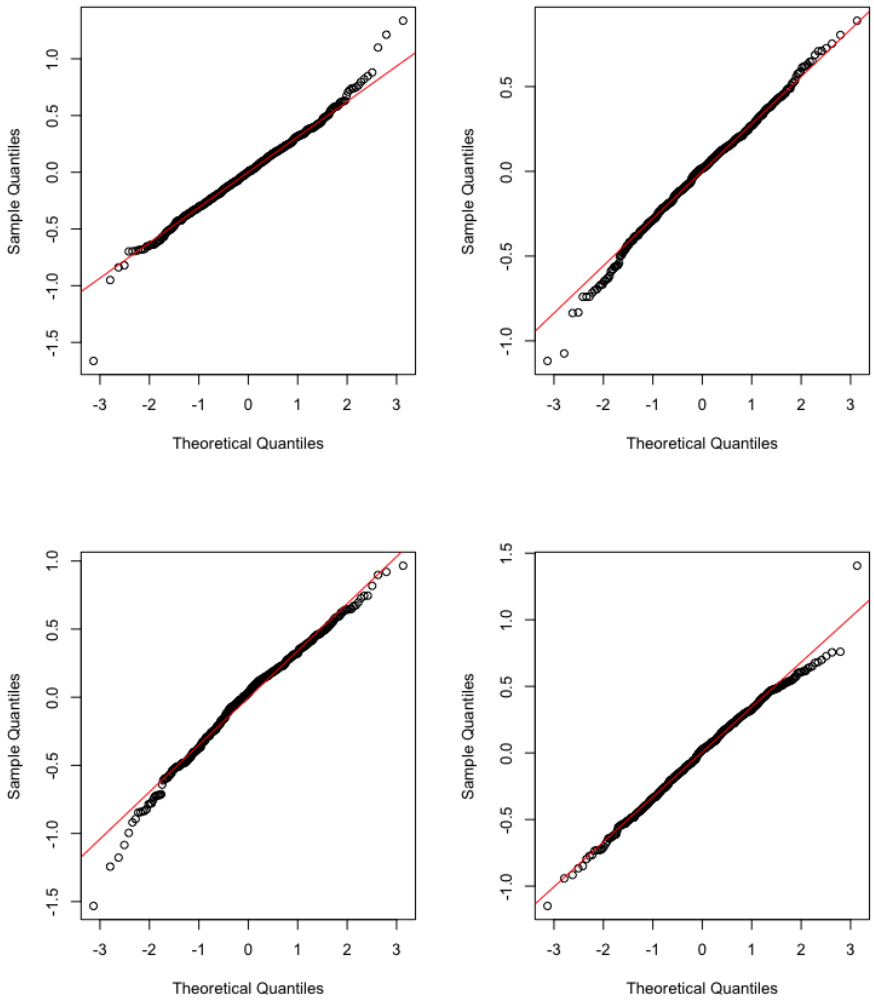

Figure 1: Normal QQ-plot of the Fully Observed Metabolites in Copsac Data

For latent factors, we assume the that $c_1, \ldots, c_n$ are independent and are independent of $\{e_{gi}\}_{g\in[p],i\in[n]}$. But for the overall sample data, We do not assume an specific probability form for y_{gi}, due to our current ignorance of the missing mechanism.

We next introduce the missing mechanism model. Let r_i be an indicator for whether or not y_i is observed, ($r_i = 1$ if y_i is observed and $r_i = 0$ otherwise). Missing mechanism is then

modeled as below where $\Psi()$ denotes for cumulative distribution function for t-distribution.

$$\mathbb{P}\left(r_{gi} = 1 \mid y_{gi}\right) = \Psi\left\{\alpha_g\left(y_{gi} - \delta_g\right)\right\}, \quad g \in [p]; i \in [n] \tag{2}$$

As you may see in the above equation, there are two metabolites-dependent parameters. α_g and $\delta_g \in \mathbb{R}$ are scale and location parameter, respectively. They behave in a way that $\alpha_g \searrow 0$ implies that the metabolites are missing completely at random (MCAR), and $\alpha_g \nearrow \infty$ implies y_{gi} is left censored at δ_g. Scale parameter δ_g, respectively.

2 is a classic model for missing data in untargeted mass spectrometry experiments[1]. Typical choices for Φ include the logistic function, an exponential probabilistic model. Here, we inherited $\Phi(x) = F_4(x)$ from our previous work([4]), which we believe is a more robust option. This has previously been used as a robust alternative to logistic and probit functions.

We assume $(y_1, r_1), ..., (y_n, r_n)$ are independent and identical to each other. Conditional on $\boldsymbol{Y} = \{y_{gi}\}_{g\in[p],i\in[n]} \in \mathbb{R}^{p\times n}$, $\{r_{gi}\}_{g\in[p],i\in[n]}$ is independent.

3.0 Estimation and Inference of Coefficients of Phenotype

In this chapter, I will introduce how to get better estimates for the coefficients when there are data missing not at random(MNAR). Then I will give an estimation of the coefficients variance. In the end, I will compare our method to other existing widely-used methods, like imputation. It will be showed that compared to classic imputation, our method gives an overall more precise and robust estimations.

3.1 Efficient Estimation by MLE

Maximum likelihood estimation(MLE) is one of the most widely used method in modern statistics to estimate the parameters of a probability distribution. By maximizing a likelihood function over feasible parameter space, the observed data is most probable under the assumed statistical model.

Considering that estimating latent factors is not the focus of this thesis, I will treat the latent factors C as known terms in the rest of this thesis. In this case, I could develop a better estimate of coefficients for both observed and latent factors based on our original estimates. For simplicity, I will rewrite the mean model for the remainder of this chapter as:

$$y_{gi} = \boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{\eta}_g + e_{gi}, \quad e_{gi} \sim (0, \sigma_g^2), \quad g \in [p]; i \in [n] \tag{3}$$

where $\boldsymbol{z}_i$ represents observed and latent factors, which are assumed to be fully observed here.

Based on our normal assumption and the missing mechanism model, we can write the log-likelihood function for this model, up to constants that do not depend on η, σ, α and δ,

is:

$$
\begin{aligned}
& l\{(\boldsymbol{y}_1, r_1) \ldots, (\boldsymbol{y}_n, r_n) \mid \boldsymbol{\eta}, \sigma, \alpha, \delta\} \\
=& \log \left\{ \prod_{i=1}^{n} \left\{ \mathcal{N}(\boldsymbol{y}_i; \boldsymbol{z}_i^T \boldsymbol{\eta}, \sigma^2) \mathbf{P}\left(r_i = 1 \mid \boldsymbol{y}_i\right) \right\}^{r_i} \mathbf{P}(r_i = 0 \mid \boldsymbol{z}_i)^{1-r_i} \right\} \\
\propto & \sum_{i=1}^{n} r_i \left\{ -\frac{1}{2\sigma^2} (\boldsymbol{y}_i - \boldsymbol{z}_i^T \boldsymbol{\eta})^2 + \log \mathbf{P}(r_i = 1 \mid \boldsymbol{y}_i) \right\} + \sum_{i=1}^{n} (1 - r_i) \log \mathbf{P}(r_i = 0 \mid \boldsymbol{z}_i)
\end{aligned} \tag{4}
$$

where $\mathbf{P}\left(r_i = 1 \mid \boldsymbol{y}_i\right) = \Psi\{\alpha(\boldsymbol{y}_i - \delta)\}$ depends on α and δ, while $\mathbf{P}\left(r_i = 0 \mid \boldsymbol{z}_i\right)$ depends on η, σ, α and δ,

$$
\begin{aligned}
\mathbf{P}(\boldsymbol{r}_i = 0 \mid \boldsymbol{z}_i) = 1 - \mathbf{P}(\boldsymbol{r}_i = 1 \mid \boldsymbol{z}_i) &= 1 - \int \mathbf{P}(\boldsymbol{r}_i = 1 \mid \boldsymbol{y}_i) \mathbf{P}(\boldsymbol{y}_i \mid \boldsymbol{z}_i) d\boldsymbol{y}_i \\
&= 1 - \int \Psi\left\{\alpha(\boldsymbol{z}_i^T \boldsymbol{\eta} + \sigma\epsilon - \delta)\right\} \phi(\epsilon) d\epsilon
\end{aligned}
$$

Let $f(x) = \frac{d\Psi(x)}{dx}$ and $\dot{f}(x) = \frac{df(x)}{dx}$, and $\phi(x)$ be the pdf for the standard normal distribution. We then could define the following functions for the sake of convenience

$$
\begin{aligned}
g_{1,i}(\boldsymbol{\eta}) &= \int \Psi\left\{\alpha(\boldsymbol{z}_i^T \boldsymbol{\eta} + \sigma\epsilon - \delta)\right\} \phi(\epsilon) d\epsilon, \\
g_{2,i}(\boldsymbol{\eta}) &= \int f\left\{\alpha(\boldsymbol{z}_i^T \boldsymbol{\eta} + \sigma\epsilon - \delta)\right\} \phi(\epsilon) d\epsilon, \\
g_{3,i}(\boldsymbol{\eta}) &= \int \dot{f}\left\{\alpha(\boldsymbol{z}_i^T \boldsymbol{\eta} + \sigma\epsilon - \delta)\right\} \phi(\epsilon) d\epsilon.
\end{aligned}
$$

Since this thesis is not focused on inference about missing mechanism, I assume for simplicity that α,δ,z_i and σ^2 are known. Detailed derivation about the parameters could be found at [4]. Thus up to constants that do not depend on $\boldsymbol{\eta}$, we can write the log-likelihood, score function and Hessian as

$$
l(\boldsymbol{\eta}) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} r_i (\boldsymbol{y}_i - \boldsymbol{z}_i^T \boldsymbol{\eta})^2 + \sum_{i=1}^{n} (1 - r_i) \log\left\{1 - g_{1,i}(\boldsymbol{\eta})\right\} \in \mathbb{R} \tag{5a}
$$

$$
\boldsymbol{s}(\boldsymbol{\eta}) = \nabla_\eta l(\boldsymbol{\eta}) = \sum_{i=1}^{n} \left\{ \frac{r_i}{\sigma^2} \left(y_i - \boldsymbol{z}_i^T \boldsymbol{\eta}\right) - (1 - r_i) \frac{\alpha g_{2,i}(\boldsymbol{\eta})}{1 - g_{1,i}(\boldsymbol{\eta})} \right\} \boldsymbol{z}_i \in \mathbb{R}^d \tag{5b}
$$

$$
\boldsymbol{H}(\boldsymbol{\eta}^T) = \nabla_\eta \boldsymbol{s}(\boldsymbol{\eta}) = -\sum_{i=1}^{n} \left\{ \frac{r_i}{\sigma^2} + (1 - r_i)\alpha^2 \left[\frac{g_{3,i}(\boldsymbol{\eta})}{1 - g_{1,i}(\boldsymbol{\eta})} + \left(\frac{g_{2,i}(\boldsymbol{\eta})}{1 - g_{1,i}(\boldsymbol{\eta})} \right)^2 \right] \right\} \boldsymbol{z}_i \boldsymbol{z}_i^T \in \mathbb{R}^{d \times d} \tag{5c}
$$

Then we can get our estimation by numerically optimizing the likelihood function for metabolites with missing proportion between 5% to 50%. As I've said above, this thesis aims to develop an efficient and easy-to-use software for the furture metabolomics research. Our algorithm are designed to be both accurate and computing-efficient. There are multiple ways to do numerical optimization in high dimension, in which we found BFGS algorithm is the most suitable method for us. Given a fair starting value from the MetabMiss software, we managed to keep the run-time for one simulation in under 5 seconds.

3.2 Finite Sample-Corrected Variance of Coefficients Estimations

In this section, I'll introduce how we get a good estimate for the variance of coefficients estimator with sample correction. To get variance of coefficients estimations, which is denoted as $\mathbb{V}(\hat{\boldsymbol{\eta}})$, we first have, by Taylor's Theorem,

$$\hat{\boldsymbol{\eta}} - \boldsymbol{\eta} = \boldsymbol{L}\left(\tilde{\boldsymbol{\eta}}\right) \boldsymbol{s}\left(\boldsymbol{\eta}\right),$$

where $\tilde{\boldsymbol{\eta}}$ lies on the line adjoining $\boldsymbol{\eta}$ and $\hat{\boldsymbol{\eta}}$ and $\boldsymbol{L}\left(\tilde{\boldsymbol{\eta}}\right) = \left\{-\boldsymbol{H}\left(\tilde{\boldsymbol{\eta}}\right)\right\}^{-1}$. Compare to $\boldsymbol{s}\left(\boldsymbol{\eta}\right)$, where the uncertainty mainly from, we temporarily ignore the uncertainty in $\boldsymbol{L} = \boldsymbol{L}\left(\tilde{\boldsymbol{\eta}}\right)$, and treat it as a known constant, then we can write

$$\mathbb{V}\left(\hat{\boldsymbol{\eta}}\right) = \mathbb{V}\left\{\boldsymbol{L}\left(\tilde{\boldsymbol{\eta}}\right) \boldsymbol{s}\left(\boldsymbol{\eta}\right)\right\} = \boldsymbol{L}(\hat{\boldsymbol{\eta}})\mathbb{V}\left\{\boldsymbol{s}\left(\boldsymbol{\eta}\right)\right\} \boldsymbol{L}(\hat{\boldsymbol{\eta}}).$$

However, when the dimension of coefficients increases, the estimation of variance will be significantly deflated. This phenomenon widely exists when estimating high dimensional coefficients due to the large degrees of freedom they brings into the system. In OLS, we have a mutual inflation factor to solve this problem by multiplying the estimation to a factor in the form of $\frac{1}{n-p}$. But when there exist missing values, we need to develop a new way to do this. To get a finite-sample corrected estimate for $\mathbb{V}\left(\hat{\boldsymbol{\eta}}\right)$ that corrects the observed variance deflation, a better estimate for $\mathbb{V}\left\{\boldsymbol{s}\left(\boldsymbol{\eta}\right)\right\}$ is rather crucial using the data we have available.

To start, note that

$$\mathbb{V}\left\{\boldsymbol{s}\left(\boldsymbol{\eta}\right)\right\} = \sum_{i=1}^{n} \mathbb{E}\left\{\boldsymbol{s}_i\left(\boldsymbol{\eta}\right)\boldsymbol{s}_i\left(\boldsymbol{\eta}\right)^T\right\} \approx \sum_{i=1}^{n} \boldsymbol{s}_i\left(\boldsymbol{\eta}\right)\boldsymbol{s}_i\left(\boldsymbol{\eta}\right)^T,$$

where when dimension of coefficients increases, $\mathbb{E}\left\{\boldsymbol{s}_i\left(\boldsymbol{\eta}\right)\boldsymbol{s}_i\left(\boldsymbol{\eta}\right)^T\right\}$ become more and more close to zero. Thus we approximate $\mathbb{E}\left\{\boldsymbol{s}_i\left(\boldsymbol{\eta}\right)\boldsymbol{s}_i\left(\boldsymbol{\eta}\right)^T\right\}$ with the third term in the above equation to extend the variance inflation factor developed in [4] to these data. As we defined in 5, for each $i = 1, \ldots, n$, we could write the score function as the following:

$$\boldsymbol{s}_i\left(\boldsymbol{\eta}\right) = \left\{\frac{r_i}{\sigma^2}\left(y_i - \boldsymbol{z}_i^T\boldsymbol{\eta}\right) - \alpha\left(1 - r_i\right)h_i\left(\boldsymbol{\eta}\right)\right\}\boldsymbol{z}_i,$$

where

$$\begin{aligned} h_i\left(\boldsymbol{\eta}\right) &= \frac{g_{2,i}(\boldsymbol{\eta})}{1 - g_{1,i}(\boldsymbol{\eta})} \\ \dot{h}_i\left(\boldsymbol{\eta}\right) &= \alpha\left[\frac{g_{3,i}\left(\boldsymbol{\eta}\right)}{1 - g_{1,i}\left(\boldsymbol{\eta}\right)} + \left\{\frac{g_{2,i}\left(\boldsymbol{\eta}\right)}{1 - g_{1,i}\left(\boldsymbol{\eta}\right)}\right\}^2\right]. \end{aligned}$$

Considering the two different situations(observed and unobserved), we split estimating score functions $\boldsymbol{s}_i(\boldsymbol{\eta})$ into two parts(detailed prove and explanation could be found at Appendix):

1. $\underline{r_i = 1}$: When the metabolites are observed, we can utilize the information from observed data point.

$$\begin{aligned} \boldsymbol{s}_i\left(\hat{\boldsymbol{\eta}}\right) &= \frac{r_i}{\sigma^2}\underbrace{\left(y_i - \boldsymbol{z}_i^T\hat{\boldsymbol{\eta}}\right)}_{\hat{\epsilon}_i}\boldsymbol{z}_i = \frac{r_i}{\sigma^2}\left\{\epsilon_i - \boldsymbol{z}_i^T\boldsymbol{L}\left(\tilde{\boldsymbol{\eta}}\right)\boldsymbol{s}\left(\boldsymbol{\eta}\right)\right\}\boldsymbol{z}_i \\ &= \frac{r_i}{\sigma^2}\left[\epsilon_i\left\{1 - \frac{r_i}{\sigma^2}\boldsymbol{z}_i^T\boldsymbol{L}\left(\tilde{\boldsymbol{\eta}}\right)\boldsymbol{z}_i\right\} - \boldsymbol{z}_i^T\boldsymbol{L}\left(\tilde{\boldsymbol{\eta}}\right)\sum_{j\neq i}\boldsymbol{s}_j\left(\boldsymbol{\eta}\right)\right]\boldsymbol{z}_i, \end{aligned}$$

where $\mathbb{E}\left\{\boldsymbol{z}_i^T\boldsymbol{L}\left(\tilde{\boldsymbol{\eta}}\right)\sum_{j\neq i}\boldsymbol{s}_j\left(\boldsymbol{\eta}\right) \mid \epsilon_i, r_i\right\} = \mathbf{0}$, this expectation is justified under the assumption that we could ignore the uncertainty of $\boldsymbol{L}\left(\tilde{\boldsymbol{\eta}}\right)$. Then it will give us the inflated estimation:

$$\boldsymbol{s}_i\left(\boldsymbol{\eta}\right)\boldsymbol{s}_i\left(\boldsymbol{\eta}\right)^T \approx \left\{1 - \frac{r_i}{\sigma^2}\boldsymbol{z}_i^T\boldsymbol{L}\left(\hat{\boldsymbol{\eta}}\right)\boldsymbol{z}_i\right\}^{-2}\boldsymbol{s}_i\left(\hat{\boldsymbol{\eta}}\right)\boldsymbol{s}_i\left(\hat{\boldsymbol{\eta}}\right)^T.$$

2. $\underline{r_i = 0}$: When the data point is missing, it's even harder for us since there is no information from data. We turn to use Taylor expansion to the score function.

   For some $\bar{\boldsymbol{\eta}}$ between $\boldsymbol{\eta}$ and $\hat{\boldsymbol{\eta}}$,

$$\begin{aligned}
\boldsymbol{s}_i(\hat{\boldsymbol{\eta}}) &= -\alpha(1-r_i)\, h_i(\hat{\boldsymbol{\eta}})\, \boldsymbol{z}_i \\
&= -\alpha(1-r_i)\left\{h_i(\boldsymbol{\eta}) + \dot{h}_i(\bar{\boldsymbol{\eta}})\, \boldsymbol{z}_i^T(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})\right\}\boldsymbol{z}_i \\
&= -\alpha(1-r_i)\left\{h_i(\boldsymbol{\eta}) + \dot{h}_i(\bar{\boldsymbol{\eta}})\, \boldsymbol{z}_i^T \boldsymbol{L}(\tilde{\boldsymbol{\eta}})\, \boldsymbol{s}(\boldsymbol{\eta})\right\}\boldsymbol{z}_i \\
&= \boldsymbol{s}_i(\boldsymbol{\eta})\left\{1 - \alpha(1-r_i)\dot{h}_i(\bar{\boldsymbol{\eta}})\, \boldsymbol{z}_i^T \boldsymbol{L}(\tilde{\boldsymbol{\eta}})\, \boldsymbol{z}_i\right\}
\end{aligned}$$

   which gives us:

$$\boldsymbol{s}_i(\boldsymbol{\eta})\, \boldsymbol{s}_i(\boldsymbol{\eta})^\top \approx \left\{1 - (1-r_i)\alpha\dot{h}_i(\hat{\boldsymbol{\eta}})\, \boldsymbol{z}_i^T \boldsymbol{L}(\hat{\boldsymbol{\eta}})\, \boldsymbol{z}_i\right\}^{-2} \boldsymbol{s}_i(\hat{\boldsymbol{\eta}})\, \boldsymbol{s}_i(\hat{\boldsymbol{\eta}})^T.$$

Putting this all together, we get the estimation for variance of score function:

$$\begin{aligned}
\sum_{i=1}^n \boldsymbol{s}_i(\boldsymbol{\eta})\, \boldsymbol{s}_i(\boldsymbol{\eta})^T &\approx \sum_{i=1}^n \frac{1}{v_i^2} \boldsymbol{s}_i(\hat{\boldsymbol{\eta}})\, \boldsymbol{s}_i(\hat{\boldsymbol{\eta}})^T \\
v_i &= 1 - \boldsymbol{z}_i^T \boldsymbol{L}(\hat{\boldsymbol{\eta}})\, \boldsymbol{z}_i \left\{\frac{r_i}{\sigma^2} + (1-r_i)\alpha\dot{h}_i(\hat{\boldsymbol{\eta}})\right\},
\end{aligned}$$

meaning the variance-inflated estimator for $\mathbb{V}(\hat{\boldsymbol{\eta}})$ is

$$\hat{\mathbb{V}}(\hat{\boldsymbol{\eta}}) = \{\boldsymbol{H}(\hat{\boldsymbol{\eta}})\}^{-1} \left\{\sum_{i=1}^n \frac{1}{v_i^2} \boldsymbol{s}_i(\hat{\boldsymbol{\eta}})\, \boldsymbol{s}_i(\hat{\boldsymbol{\eta}})^T\right\} \{\boldsymbol{H}(\hat{\boldsymbol{\eta}})\}^{-1} \tag{6a}$$

$$v_i = 1 + \boldsymbol{z}_i^T \{\boldsymbol{H}(\hat{\boldsymbol{\eta}})\}^{-1} \boldsymbol{z}_i \left\{\frac{r_i}{\sigma^2} + (1-r_i)\alpha\dot{h}_i(\hat{\boldsymbol{\eta}})\right\} \tag{6b}$$

$$\dot{h}_i(\hat{\boldsymbol{\eta}}) = \alpha\left[\frac{g_{3,i}(\hat{\boldsymbol{\eta}})}{1 - g_{1,i}(\hat{\boldsymbol{\eta}})} + \left\{\frac{g_{2,i}(\hat{\boldsymbol{\eta}})}{1 - g_{1,i}(\hat{\boldsymbol{\eta}})}\right\}^2\right] \tag{6c}$$

which is equivalent to

$$v_i = 1 + \boldsymbol{z}_i^T \{\boldsymbol{H}(\hat{\boldsymbol{\eta}})\}^{-1} \boldsymbol{z}_i \left\{\frac{r_i}{\sigma^2} + (1-r_i)\alpha^2\left[\frac{g_{3,i}(\hat{\boldsymbol{\eta}})}{1 - g_{1,i}(\hat{\boldsymbol{\eta}})} + \left\{\frac{g_{2,i}(\hat{\boldsymbol{\eta}})}{1 - g_{1,i}(\hat{\boldsymbol{\eta}})}\right\}^2\right]\right\} \tag{7a}$$

Notice that we need to assure that $\boldsymbol{H}(\hat{\boldsymbol{\eta}})$ to be negative definite in order for the estimation procedure normally behaves.

After our experiments on simulations, we found that $\dot{h}_i(\boldsymbol{\eta})$ can sometimes be negative, which then may cause the Hessian matrix to be positive definite instead of negative definite.

Thus, we instead take expectation of the variance estimator function with respect to missing indices r_i. This is effectively replace Hessian matrix with its expectation, which is always positive. The inflated variance then become

$$\hat{\mathbb{V}}(\hat{\boldsymbol{\eta}}) = \left[\mathbb{E}\left\{\boldsymbol{H}(\hat{\boldsymbol{\eta}})\right\}\right]^{-1} \left\{\sum_{i=1}^{n} \frac{1}{\tilde{v}_i^2} \boldsymbol{s}_i(\hat{\boldsymbol{\eta}}) \boldsymbol{s}_i(\hat{\boldsymbol{\eta}})^T\right\} \left[\mathbb{E}\left\{\boldsymbol{H}(\hat{\boldsymbol{\eta}})\right\}\right]^{-1},$$

$$\tilde{v}_i = 1 + a_i \boldsymbol{z}_i^T \left[\mathbb{E}\left\{\boldsymbol{H}(\hat{\boldsymbol{\eta}})\right\}\right]^{-1} \boldsymbol{z}_i,$$

$$a_i = \frac{g_{1,i}(\boldsymbol{\eta})}{\sigma^2} + \alpha^2 \left[g_{3,i}(\boldsymbol{\eta}) + \frac{g_{2,i}(\boldsymbol{\eta})^2}{1 - g_{1,i}(\boldsymbol{\eta})}\right].$$

with the expectation of hessian being

$$\mathbb{E}\left\{\boldsymbol{H}(\boldsymbol{\eta})\right\} = -\sum_{i=1}^{n} \left(\frac{g_{1,i}(\boldsymbol{\eta})}{\sigma^2} + \alpha^2 \left[g_{3,i}(\boldsymbol{\eta}) + \frac{g_{2,i}(\boldsymbol{\eta})^2}{1 - g_{1,i}(\boldsymbol{\eta})}\right]\right) \boldsymbol{z}_i \boldsymbol{z}_i^T = -\sum_{i=1}^{n} a_i \boldsymbol{z}_i \boldsymbol{z}_i^T$$

3.3 Comparison to Other Methods in Simulation

In Mass-Spectrometry(MS) metabolites data, missing values always exist due to multiple reasons. Based on the properties of MS, many algorithms have been developed or adopted to solve this problem. In this thesis, I thoroughly compared some selected methods to our method, and the results will be displayed in each chapter, respectively. Let's begin with introduction of these methods.

3.3.1 Introduction to Imputation Methods

Imputation methods are a group of methods that are comprehensively well-developed dedicated to different assumption with respect to different intuition and experiment condition. Their intention seems intuitive yet powerful. For general MS metabolomics data, I selected five of the best-performed methods according to [2].

1. *min*: Impute missing values with the minimum value among each metabolites across samples.

   This method comes from an intuition that the missing values exist due to the true target metabolite concentration is too below to be detected in experiment. Then it is reasonable to assume the lowest metabolites concentration is approximately equal to the lowest threshold. It generally shows inflated type 1 error and low power for regression analysis[5].

2. *svd*: SVD imputation obtains a set of mutually orthogonal expression patterns that can be further linearly combined to approximate certain mechanism[6]. Later we will see SVD also appears in dimension reduction. A simplified algorithm procedure should look like below:

   a. Replace all missing values with row(each metabolites) means.
   b. Compute a rank-k(k is set to be the same as latent factors dimension) approximation to the imputed matrix.
   c. Replace values in the imputed positions with corresponding values from the approximation computed in Step 2.
   d. Repeat Steps 2 and 3 until convergence.

3. *knn-var*: k-Nearest Neighbors per metabolites.

   KNN finds k-nearest neighborhoods for each term, and use the average value among them to impute the missing value. When we choose to measure the distance between metabolites, this method shows a low type 1 error rate, but large power. especially when the amount of missing values are bigger. Thus, I finally turned to use *knn-obs-sel* in simulations.

4. *knn-obs-sel*: KNN per observation(samples) using selected metabolites.

   This method measures the distance between different samples, but using only selected part of the variables as measurement. To me, it makes more sense intuitively since samples are assumed to be independent in our model. Yet the metabolites are assumed to be correlated in form of cluster. It also shows high power and an overall marginal type 1 error rate, even for a high amount of missing values in [2].

5. *mice*: Multiple Imputation(MI) by Chained Equations.

   MI is specifically designed to improve the performance of stochastic imputation methods. Stochastic imputations normally use assumption that there exist an underlying mean model for missing values. The stochastic part is introduced to simulate the randomness when the data is generated. When applying MI, the first step is to impute the incomplete data m times to produce m complete datasets. Then statistical analysis and inference is performed on the m datasets, individually. Using tricks like averaging the total coefficients estimation, MI imputation is finalized.

   In practice, we found that MI requires much more computational resources than the previous methods, which isn't realistic in real-world metabolomics data analysis. Thus, we didn't include it in our final simulations.

3.3.2 Simulation Results

Before we start to do results analysis, it is worth mentioning that we are also comparing our method to OLS, dSVA, and method directly from MetabMiss[4], which in practice is the starting value of out current one. Detailed introduction of these methods won't be covered in this thesis since they are relatively trivial.

Noted that the basic simulation framework for this thesis is universal. We set $\boldsymbol{Z}$ to be and all-one vector corresponding to nuisance parameters, $\boldsymbol{X}$ to be a half zero half one vector corresponding to parameters of interest. For latent factors $\boldsymbol{C}$,the dimension is set to be 10. Data matrix is set to be a 1200(metabolites) by 600(samples) matrix, among which there are around 20% are fully observed; 20% are missing less than 5%, which are considered as missing at random(MAR); 25% are missing at 5%-50%, which are considered as missing not at random(MNAR).

To begin with, we consider the confidence interval coverage for the effects of interest β_g for each method in Figure 2. It shows the fraction of effects of interest $\{\beta_g\}_{g\in M}$ in all 50 simulated data sets that lie in their respective 95% confidence intervals $\hat{\boldsymbol{\beta}}_g \pm 1.96 \left\{\hat{\mathbb{V}}(\hat{\boldsymbol{\beta}}_g)\right\}^{1/2}$, stratified by the simulated "true" $|\boldsymbol{\beta}_g|$. As you can see in Figure 2, our methods universally keeps the CI coverage closest to 95%. Also, note that *min* performs relatively better than

other imputation methods since this assumption are heavily dependent to real experiment experience.

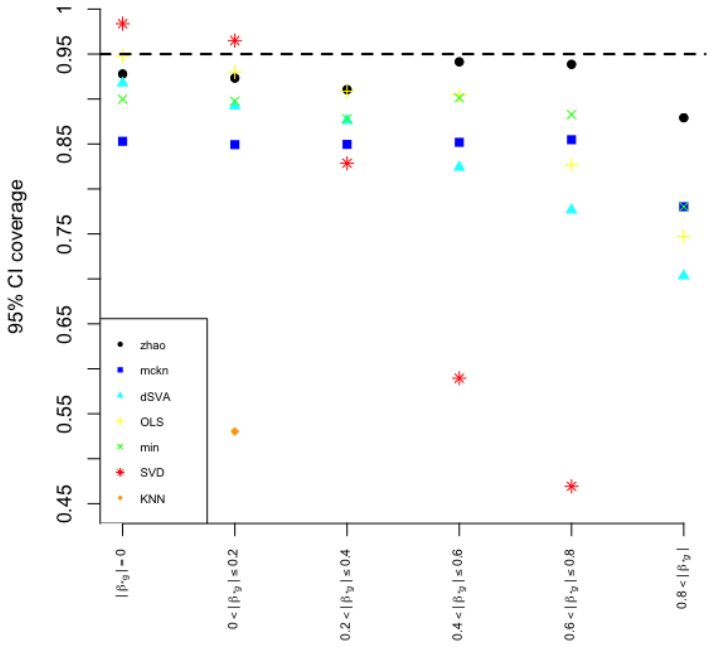

Figure 2: Confidence Interval(CI) Coverage

Figure 3, from top to bottom, from left to right, shows the false discovery proportion(FDP) and true recovery proportion(TRP) for metabolites with q-values $\leq$ 0.05, 0.1, 0.15 and 0.2.

The TRP is the fraction of metabolites with non-zero $\boldsymbol{\beta}_g$ identified at a given q-value threshold. q-values were determined using the qvalue package in R (Storey et al., 2015).

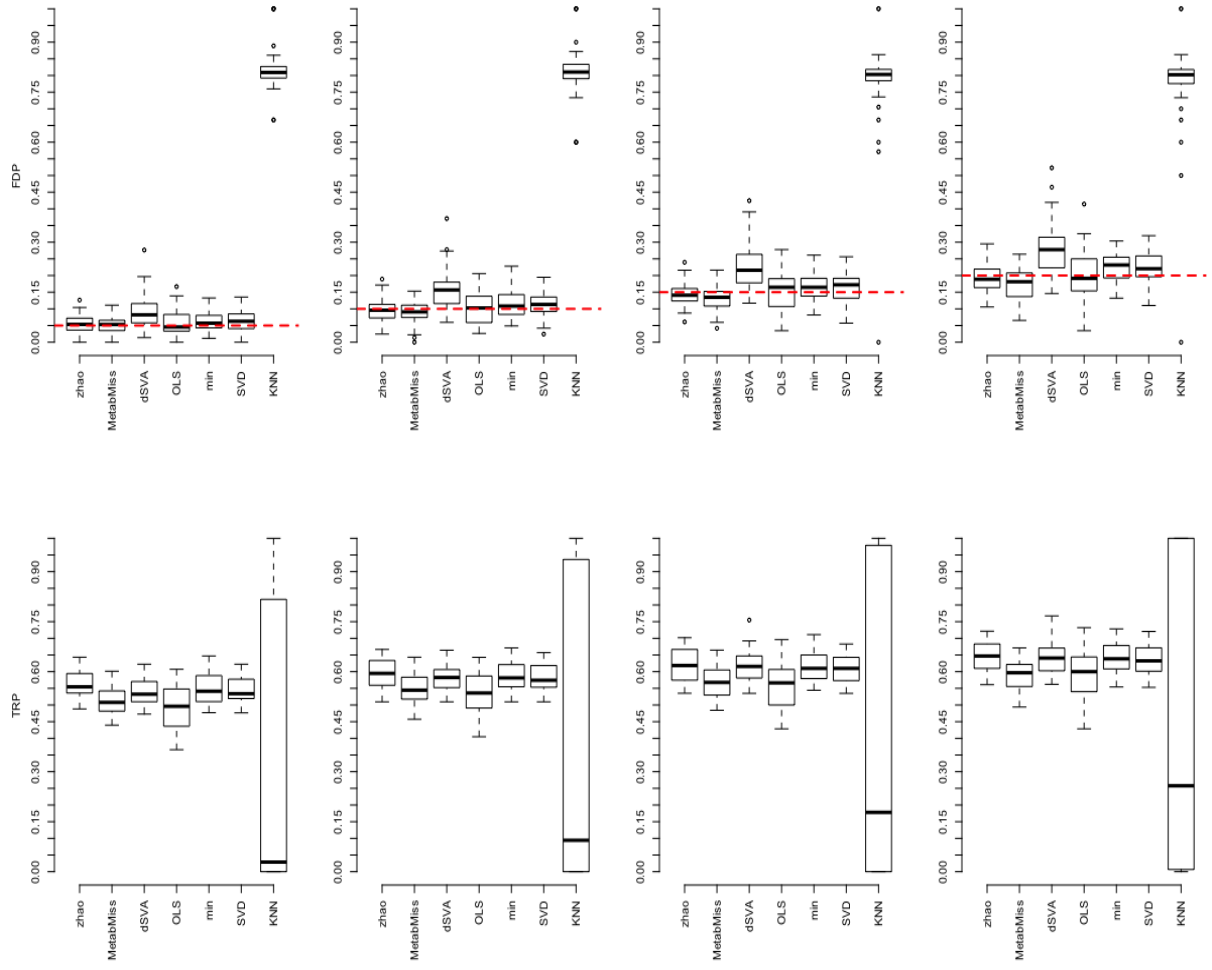

Figure 3: False Discovery Rate and True Recovery Proportion

4.0 Dimension Reduction Using Principal Component Regression

In this chapter, we will briefly introduce PCR, the commonly used dimension reduction methods. Then it will be discussed how PCR is used to help us do dimension reduction of metabolomics data. Finally I will compare our method to other algorithms mentioned in precious chapter.

4.1 Introduction to PCR

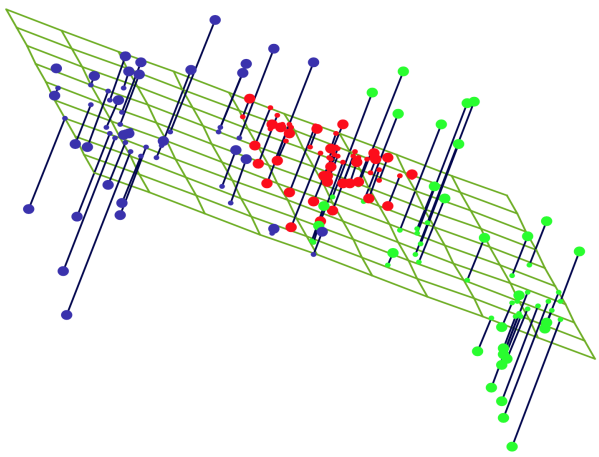

Figure 4: Principle Component Analysis from Hastie, Tibshirani and Friedman

Principal component regression (PCR) is a regression analysis technique that is based on principal component analysis (PCA), which is basically a projection computing algorithms that could be used to change basis on the coefficients of interest. More specifically, PCR is used for a high dimensional standard linear regression model. This technique could be used to explore data analysis and prediction. The plot above shows how PCA reduce the dimension where its variance is biggest when the dimension is relatively small. For linear regression model, PCR is fundamentally selecting the coefficients that can explain the variance mostly.

For a general PCA problem, we demonstrate it to be an optimization problem, where we need to first write out the general mean model:

$$f(\lambda) = \mu + \mathbf{V}_q \lambda$$

where μ is the location parameter, $\mathbf{V}_q$ here is the orthogonal basis that represent the hyperplane that can mostly explain variance and λ is a q-vector of parameters. They combined to create an hyperplane of rank-q,which is the dimension reduced representation of original data matrix. The objective function should be built such that the least squares amounts is minimized in terms of the reconstruction error[3]:

$$\mathbf{V}_q = \arg \min_{\mu, \{\lambda_i\}, \mathbf{V}_q} \sum_{i=1}^{N} \|x_i - \mu - \mathbf{V}_q \lambda_i\|^2$$

By trivial mathematical derivation, the solution is analytically expressed as follow:

1. First construct a singular value decomposition to $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$.
2. Then the first K columns of $\mathbf{UD}$ are the K principal components of $\mathbf{X}$.

4.2 Factor Analysis Using PCR

In the remainder of this chapter, we will rewrite our mean model so that it will be easier to interpret.

$$\boldsymbol{Y} = \boldsymbol{\Gamma}\boldsymbol{Z}^T + \boldsymbol{L}\boldsymbol{C}^T + \boldsymbol{E} \tag{8}$$

where $\boldsymbol{Z} \in \mathbb{R}^{n \times r}$ are observed nuisance covariates (like the intercept), which we assume $\boldsymbol{Z}^T\boldsymbol{C} = \mathbf{0}$ for identification purposes (i.e. $\boldsymbol{C}$ is orthogonal to $\boldsymbol{Z}$). For the residuals, we will also assume

$$\boldsymbol{E} \sim MN_{p \times n}(\mathbf{0}, \boldsymbol{\Sigma}, I_n),$$

where $\boldsymbol{\Sigma} = (\sigma_{gh})_{g \in [p]; h \in [p]}$ is a sparse matrix.

The goal is to do factor analysis for $\boldsymbol{L} \in \mathbb{R}^{p \times K}$. Based on the classical PCA algorithms, we adopt it to our models as follow:

1. Use MetabMiss[4] to estimate α_g, δ_g for missingness mechanism, as well as $\boldsymbol{C}, \boldsymbol{L}, \sigma_{gg}$, for all $g \in [p]$. Let $\hat{\boldsymbol{C}}, \hat{\boldsymbol{L}}$ be the estimate for $\boldsymbol{C}, \boldsymbol{L}$, respectively.
2. Use our new method to update estimation of coefficients of interest $\boldsymbol{L}$, which we do by regressing $\boldsymbol{Y}_{g\cdot}$ onto $[\boldsymbol{Z}, \hat{\boldsymbol{C}}]$ for each g with missing proportion less than 50%.
3. Under the true model, we know:

$$\mathbb{E}\left(\frac{1}{n}\boldsymbol{C}^\top\boldsymbol{C}\right) = \frac{1}{n}\sum_i \mathbb{E}(\boldsymbol{C}_{i\cdot}\boldsymbol{C}_{i\cdot}^\top) = \frac{1}{n}\sum_i \mathbb{V}(\boldsymbol{C}_{i\cdot}) = \frac{1}{n}n\boldsymbol{\Omega} = \boldsymbol{\Omega}.$$

   Since from our assumed mean model, we can only identify the expectation of the multiplication $\boldsymbol{L}\boldsymbol{C}^\top$, in which case we couldn't identify each one up to rotation. Thus for identification purpose, we need to define a transformed $\tilde{\boldsymbol{C}}$ such that

$$\frac{1}{(n-r)}\tilde{\boldsymbol{C}}^\top\boldsymbol{P}_Z^\perp\tilde{\boldsymbol{C}} = \boldsymbol{I}_K.$$

   So that we could let $\tilde{\boldsymbol{C}} = \hat{\boldsymbol{C}}\hat{\boldsymbol{R}}^{-1}$ where $\hat{\boldsymbol{R}} \in \mathbb{R}^{K\times K}$ be a symmetric matrix that satisfies

$$\frac{1}{(n-r)}\hat{\boldsymbol{C}}^\top\hat{\boldsymbol{C}} = \hat{\boldsymbol{R}}^2.$$

4. To guarantee $\mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{Z}) = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{C}}^\top = \hat{\boldsymbol{L}}\hat{\boldsymbol{C}}^\top$, we also let $\tilde{\boldsymbol{L}} = \hat{\boldsymbol{L}}\hat{\boldsymbol{R}}$.
5. Then we do eigenvalue decomposition to $\tilde{\boldsymbol{L}}^\top\tilde{\boldsymbol{L}}$ as follow

$$\frac{(n-r)}{p}\tilde{\boldsymbol{L}}^\top\tilde{\boldsymbol{L}} = \frac{(n-r)}{p}\hat{\boldsymbol{R}}\hat{\boldsymbol{L}}^\top\hat{\boldsymbol{L}}\hat{\boldsymbol{R}} = \hat{\boldsymbol{U}}\begin{pmatrix}\hat{\lambda}_1 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \hat{\lambda}_K\end{pmatrix}\hat{\boldsymbol{U}}^\top.$$

   where $\hat{\boldsymbol{U}} \in \mathbb{R}^{K\times K}$ be a unitary matrix
6. The PCA estimates for the $\boldsymbol{C}$ and $\boldsymbol{L}$ then, are $\hat{\boldsymbol{C}}^{(PCA)} = \hat{\boldsymbol{C}}\hat{\boldsymbol{R}}^{-1}\hat{\boldsymbol{U}}$ and $\hat{\boldsymbol{L}}^{(PCA)} = \hat{\boldsymbol{L}}\hat{\boldsymbol{R}}\hat{\boldsymbol{U}}$. It can be easily checked that both of these estimates have orthogonal columns.

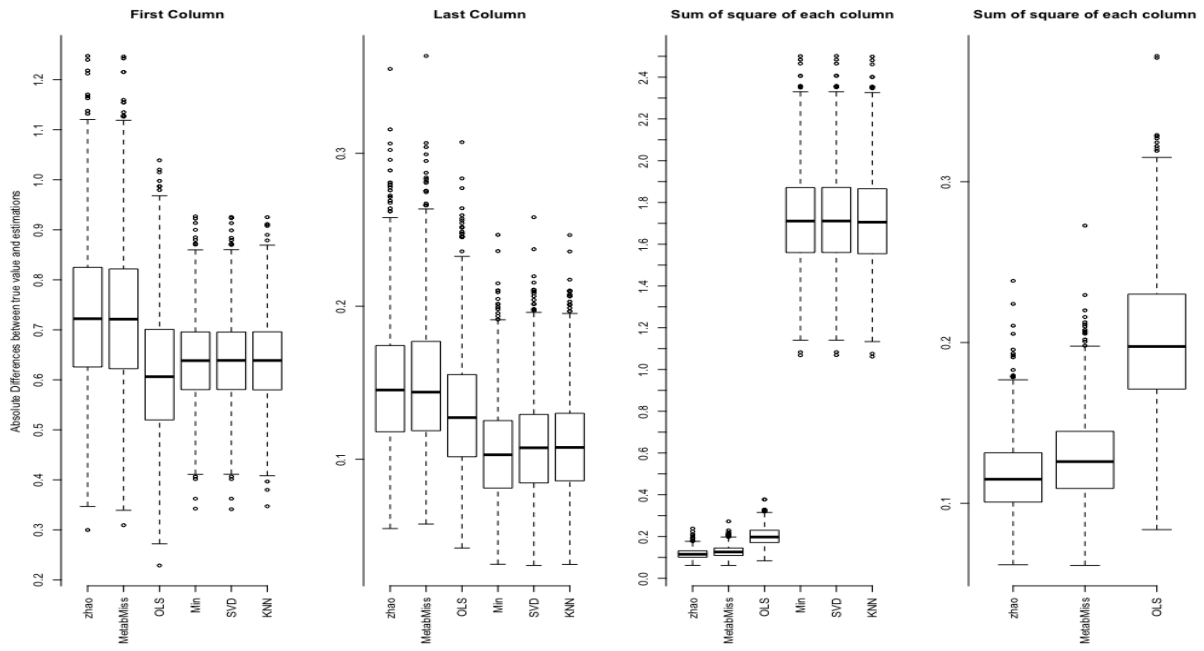

Figure 5: Principle Component Regression

4.3 Simulations

In this section, we demonstrate the result by showing the absolute difference between the PCR of simulation and PCA of other estimation methods. As for the simulation parameters, they are basically the same as the setup in chapter 3, except that we assume there are no other coefficients other than intercept and latent factors.

The first two plots in Figure 5 is about the absolute difference between estimating methods and the simulated parameters of first and last column of the PCR result. It clearly shows that our methods is doing the best job finding principle component even from the latent factors. For the last two plots, they compare row sum of square of $\hat{\boldsymbol{L}}_{PCA}$ by absolute difference,. This term gives us a proportion variance explained by C, which is identifiable. The last plot is the same as the third one with only the first three terms, which gives us a better point of view of our model's performance.

5.0 Estimation of Correlation between Metabolites

In this chapter, we will introduce how do we get the estimation for correlation coefficients between metabolites. Also, based on the estimator procedure, we can easily get the estimation of correlation covariance.

5.1 Correlation Estimator and Covariance

Similar as chapter 3,for the metabolites correlations, we could easily write the likelihood functions based on our previous analysis. For each pair of metabolites, there are total n pairs of samples, which could be further divided into three different kinds. Then we could write this problem as an optimization problem with form

$$\hat{\rho}_{gh} = \arg\max \prod_{i=1}^{N} c_i(\rho_{gh}),$$

$$c_i(\rho_{gh}) = \begin{cases} c_i^{11}(\rho_{gh}) & \text{if } y_{gi} \text{ and } y_{hi} \text{ are both observed} \\ c_i^{10}(\rho_{gh}) & \text{if one of both observed.} \\ c_i^{00}(\rho_{gh}) & \text{if non of both are observed.} \end{cases}$$

Based on the three different kinds of sample pairs, we established three different likelihood using Bayesian methodology that is given below.

If both metabolites are observed

$$\begin{aligned} c_i^{11} &= \log P(\rho_{gh} \mid r_{gi} = r_{hi} = 1, y_{gi}, y_{hi}) \\ &= \log P(r_{gi} = r_{hi} = 1 \mid y_{gi}, y_{hi}) \mathcal{N}(y_{gi}, y_{hi}) \\ &\propto \log \mathcal{N}(y_{gi}, y_{hi}) \end{aligned}$$

If only one metabolite of both is observed

$$\begin{aligned} c_i^{10} &= \log P(\rho_{gh} \mid r_{gi} = 1, r_{hi} = 0, y_{gi}) \\ &= \log P(r_{gi} = 1, r_{hi} = 0 \mid y_{gi}) P(y_{gi}) \\ &= \log \int P(r_{gi} = 1, r_{hi} = 0 \mid y_{hi}, y_{gi}) P(y_{hi} \mid y_{gi})\, dy_{hi} \\ &\propto \log \int (1 - \Psi\{\alpha_h(y_{hi} - \delta_h)\})\, \mathcal{N}(y_{hi} \mid y_{gi})\, dy_{hi} \end{aligned}$$

If both metabolites are missing

$$\begin{aligned} c_i^{00} &= \log P(\rho_{gh} \mid r_{gi} = r_{hi} = 0) \\ &= \log \iint P(r_{gi} = 0 \mid y_{gi}) P(r_{hi} = 0 \mid y_{hi}) P(y_{gi}, y_{hi})\, dy_{gi} dy_{hi} \\ &= \log \iint (1 - \Psi\{\alpha_g(y_{gi} - \delta_g)\})\, (1 - \Psi\{\alpha_h(y_{hi} - \delta_h)\})\, \mathcal{N}(y_{gi}, y_{hi})\, dy_{gi} dy_{hi} \end{aligned}$$

By solving this optimization problem numerically, then could get our estimation for correlation matrix. Detailed derivation that is used in software package will be supplied in Appendix.

Naturally, we could write the covariance estimator as

$$\mathbb{V}(\rho_{gh}) = \frac{1}{\sum_i \dot{c}_i^{11} + \dot{c}_i^{10} + \dot{c}_i^{00}}$$

where $\dot{c}_i$ denote the derivation of each likelihood function above, respectively.

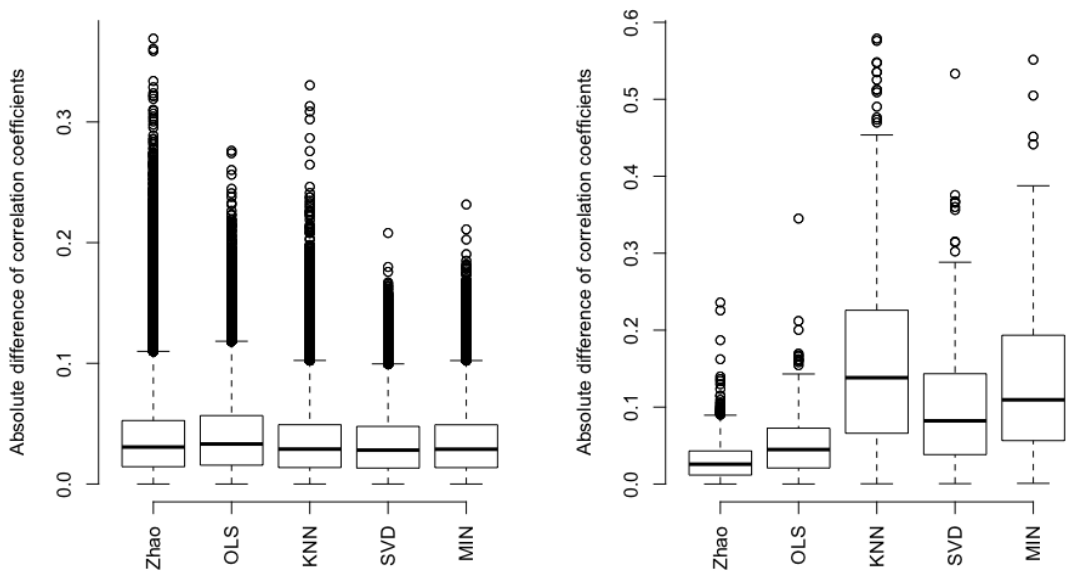

Figure 6: Correlation Estimation

5.2 Simulation

In this chapter's simulation, it's worth mentioning that we simulate our covariance matrix as a block-diagonal matrix, which means metabolites are correlated within clusters but uncorrelated outside of clusters.

From left to right, Figure 6 compares each estimations in absolute difference, where true correlation being zero and non-zero, respectively. Note that when the true values are zero, all methods performs similar, among which SVD imputation performs slightly better than the others. It makes sense because SVD is heavily relying on the assumption that the true correlation is equal to zero. But when the true value is non-zero, difference between methods starts to show, and SVD is doing much worse.

6.0 Conclusion

In this thesis, I studied the metabolomics data with non-ignorable missing data. Conditional on the missingness mechanism and latent factor estimation methods from MetabMiss [4], it is possible for us to utilize the normality assumption buried in metabolomics data and get a much powerful estimation.

I evaluated our estimation in three different chapters where each one focus on one point. By using the normality assumption, we are able to build maximal likelihood estimation model for metabolomics data. We showed that overall my method has a consistent and accurate estimate for the confident intervals. Also, we showed it has a much better performance in terms to false discovery rate control and power.

Then I showed how I am able to do dimension reduction with missing not at random data. It showed that based on our estimation method, PCA gives us a decreasing proportion estimation of variance with orthogonal regression parameters.

For the last parts, I showed the estimation of correlation between different metabolites and how my methods achieved a better performs especially when the true correlation coefficients are zero.

In my thesis, we thoroughly investigated the estimation and inference methods in biological data with missing not at random values. Future work are expected to include seeking a more fundamental proof of our normality assumption. Besides that, we may seek to develop a more powerful way to account for the missing mechanism with this normality assumptions.

Appendix A Inflated Variance Estimator

$\underline{r_i = 1}$: When the metabolites are observed, we can utilize the information from observed data point.

$$\begin{aligned}\boldsymbol{s}_i(\hat{\boldsymbol{\eta}}) &= \frac{r_i}{\sigma^2}\underbrace{\left(y_i - \boldsymbol{z}_i^T\hat{\boldsymbol{\eta}}\right)}_{\hat{\epsilon}_i}\boldsymbol{z}_i = \frac{r_i}{\sigma^2}\left\{\epsilon_i - \boldsymbol{z}_i^T\boldsymbol{L}(\tilde{\boldsymbol{\eta}})\,\boldsymbol{s}(\boldsymbol{\eta})\right\}\boldsymbol{z}_i \\ &= \frac{r_i}{\sigma^2}\left[\epsilon_i\left\{1 - \frac{r_i}{\sigma^2}\boldsymbol{z}_i^T\boldsymbol{L}(\tilde{\boldsymbol{\eta}})\,\boldsymbol{z}_i\right\} - \boldsymbol{z}_i^T\boldsymbol{L}(\tilde{\boldsymbol{\eta}})\sum_{j\neq i}\boldsymbol{s}_j(\boldsymbol{\eta})\right]\boldsymbol{z}_i,\end{aligned}$$

where $\mathbb{E}\left\{\boldsymbol{z}_i^T\boldsymbol{L}(\tilde{\boldsymbol{\eta}})\sum_{j\neq i}\boldsymbol{s}_j(\boldsymbol{\eta}) \mid \epsilon_i, r_i\right\} = \mathbf{0}$, this expectation is justified under the assumption that we could ignore the uncertainty of $\boldsymbol{L}(\tilde{\boldsymbol{\eta}})$. Then it will give us the inflated estimation:

$$\boldsymbol{s}_i(\boldsymbol{\eta})\,\boldsymbol{s}_i(\boldsymbol{\eta})^T \approx \left\{1 - \frac{r_i}{\sigma^2}\boldsymbol{z}_i^T\boldsymbol{L}(\hat{\boldsymbol{\eta}})\,\boldsymbol{z}_i\right\}^{-2}\boldsymbol{s}_i(\hat{\boldsymbol{\eta}})\,\boldsymbol{s}_i(\hat{\boldsymbol{\eta}})^T.$$

$\underline{r_i = 0}$: When the data point is missing, it's even harder for us since there is no information from data. We turn to use Taylor expansion to the score function.

For some $\bar{\boldsymbol{\eta}}$ between $\boldsymbol{\eta}$ and $\hat{\boldsymbol{\eta}}$,

$$\begin{aligned}\boldsymbol{s}_i(\hat{\boldsymbol{\eta}}) &= -\alpha(1-r_i)\,h_i(\hat{\boldsymbol{\eta}})\,\boldsymbol{z}_i \\ &= -\alpha(1-r_i)\left\{h_i(\boldsymbol{\eta}) + \dot{h}_i(\bar{\boldsymbol{\eta}})\,\boldsymbol{z}_i^T(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})\right\}\boldsymbol{z}_i \\ &= -\alpha(1-r_i)\left\{h_i(\boldsymbol{\eta}) + \dot{h}_i(\bar{\boldsymbol{\eta}})\,\boldsymbol{z}_i^T\boldsymbol{L}(\tilde{\boldsymbol{\eta}})\,\boldsymbol{s}(\boldsymbol{\eta})\right\}\boldsymbol{z}_i \\ &= \boldsymbol{s}_i(\boldsymbol{\eta})\left\{1 - \alpha(1-r_i)\dot{h}_i(\bar{\boldsymbol{\eta}})\,\boldsymbol{z}_i^T\boldsymbol{L}(\tilde{\boldsymbol{\eta}})\,\boldsymbol{z}_i\right\}\end{aligned}$$

which gives us:

$$\boldsymbol{s}_i(\boldsymbol{\eta})\,\boldsymbol{s}_i(\boldsymbol{\eta})^\top \approx \left\{1 - (1-r_i)\alpha\dot{h}_i(\hat{\boldsymbol{\eta}})\,\boldsymbol{z}_i^T\boldsymbol{L}(\hat{\boldsymbol{\eta}})\,\boldsymbol{z}_i\right\}^{-2}\boldsymbol{s}_i(\hat{\boldsymbol{\eta}})\,\boldsymbol{s}_i(\hat{\boldsymbol{\eta}})^T.$$

Appendix B Basic Derivation of Mean Model

Let $\boldsymbol{Y} \in \mathbb{R}^{p \times n}$ be the #metabolite by #sample data matrix. Some of the entries of $\boldsymbol{Y}$ are missing not at random (MNAR).

$$\boldsymbol{Y} = \boldsymbol{L}\boldsymbol{C}^\top + \boldsymbol{E}, \quad \boldsymbol{C} \in \mathbb{R}^{n \times K}, \quad \boldsymbol{L} \in \mathbb{R}^{p \times K}$$

For $\boldsymbol{C}$, we have the following properties:

$$\begin{aligned} \mathbb{E}(\boldsymbol{C}) &= 0_{n \times K} \\ \mathbb{V}(\boldsymbol{C}_{i\cdot}) &= \boldsymbol{\Omega}_{K \times K} \\ \mathbb{V}(\boldsymbol{C}_{i\cdot}, \boldsymbol{C}_{j\cdot}) &= 0_{K \times K} \end{aligned}$$

For residuals $\boldsymbol{E}$, we have the following properties:

$$\begin{aligned} \mathbb{E}(\boldsymbol{E}) &= 0_{n \times K} \\ \mathbb{V}(\boldsymbol{E}_{g\cdot}) = \sigma_g^2 I_n \quad & \mathbb{V}(\boldsymbol{E}_{g\cdot}, \boldsymbol{E}_{h\cdot}) = \sigma_{gh} \boldsymbol{I}_n \\ \mathbb{V}(\boldsymbol{E}_{\cdot i}) = \boldsymbol{\Sigma}_{p \times p} \quad & \mathbb{V}(\boldsymbol{E}_{\cdot i}, \boldsymbol{E}_{\cdot j}) = 0_{p \times p} \end{aligned}$$

We could rewrite the equation as:

$$\boldsymbol{y}_{g\cdot} = \boldsymbol{C}\boldsymbol{L}_{g\cdot} + \boldsymbol{E}_{g\cdot}$$

$$
\begin{aligned}
\mathbb{V}(\boldsymbol{y}_{\cdot i}) &= \mathbb{V}(\boldsymbol{L}\boldsymbol{C}_{i\cdot}) + \mathbb{V}(\boldsymbol{E}_{\cdot i}) \\
&= \boldsymbol{L}\mathbb{V}(\boldsymbol{C}_{i\cdot})\boldsymbol{L}^\top + \boldsymbol{\Sigma} \\
&= \boldsymbol{L}\boldsymbol{\Omega}\boldsymbol{L}^\top + \boldsymbol{\Sigma} \\
\mathbb{V}(\boldsymbol{y}_{g\cdot}) &= \mathbb{V}(\boldsymbol{C}\boldsymbol{L}_{g\cdot}) + \mathbb{V}(\boldsymbol{E}_{g\cdot}) \\
&= \mathbb{V}(\boldsymbol{C}_{i\cdot}^\top \boldsymbol{L}_{g\cdot})\boldsymbol{I}_n + \mathbb{V}(\boldsymbol{E}_{g\cdot}) \\
&= \boldsymbol{L}_{g\cdot}^\top \boldsymbol{\Omega}\boldsymbol{L}_{g\cdot}\boldsymbol{I}_n + \sigma_g^2 \boldsymbol{I}_n \\
\mathbb{V}(\boldsymbol{y}_{g\cdot}, \boldsymbol{y}_{h\cdot}) &= \mathbb{V}(\boldsymbol{C}\boldsymbol{L}_{g\cdot} + \boldsymbol{E}_{g\cdot}, \boldsymbol{C}\boldsymbol{L}_{h\cdot} + \boldsymbol{E}_{h\cdot}) \\
&= \mathbb{V}(\boldsymbol{C}\boldsymbol{L}_{g\cdot}, \boldsymbol{C}\boldsymbol{L}_{h\cdot}) + \mathbb{V}(\boldsymbol{E}_{g\cdot}, \boldsymbol{E}_{h\cdot}) + \mathbb{V}(\boldsymbol{C}\boldsymbol{L}_{g\cdot}, \boldsymbol{E}_{h\cdot}) + \mathbb{V}(\boldsymbol{C}\boldsymbol{L}_{h\cdot}, \boldsymbol{E}_{g\cdot}) \\
&= \boldsymbol{L}_{g\cdot}^\top \boldsymbol{\Omega}\boldsymbol{L}_{h\cdot}\boldsymbol{I}_n + 0 + 0 + 0 \\
&= \boldsymbol{L}_{g\cdot}^\top \boldsymbol{\Omega}\boldsymbol{L}_{h\cdot}\boldsymbol{I}_n \\
\mathbb{V}(\boldsymbol{y}_{\cdot i}, \boldsymbol{y}_{\cdot j}) &= \mathbb{V}(\boldsymbol{L}\boldsymbol{C}_{i\cdot} + \boldsymbol{E}_{\cdot i}, \boldsymbol{L}\boldsymbol{C}_{j\cdot} + \boldsymbol{E}_{\cdot j}) \\
&= \mathbb{V}(\boldsymbol{L}\boldsymbol{C}_{i\cdot}, \boldsymbol{L}\boldsymbol{C}_{j\cdot}) + \mathbb{V}(\boldsymbol{E}_{\cdot j}, \boldsymbol{E}_{\cdot j}) + \mathbb{V}(\boldsymbol{L}\boldsymbol{C}_{i\cdot}, \boldsymbol{E}_{\cdot j}) + \mathbb{V}(\boldsymbol{L}\boldsymbol{C}_{j\cdot}, \boldsymbol{E}_{\cdot i}) \\
&= \boldsymbol{L}\mathbb{V}(\boldsymbol{C}_{i\cdot}\boldsymbol{C}_{j\cdot})\boldsymbol{L}^\top \\
&= 0
\end{aligned}
$$

Appendix C Derivation of Likelihood Functions of Correlation Estimation

C.1 Estimating the Correlation between Metabolites

If both metabolites are fully observed or at least with less than 0.05 missing values

$$\begin{aligned}\log P_{r_{gi}=r_{hi}=1,y_{gi},y_{hi}}(\rho_{gh}) &= \log P(r_{gi}=r_{hi}=1 \mid y_{gi},y_{hi})\phi(y_{gi},y_{hi}) \\ &\propto \log dMVN(y_{gi},y_{hi})\end{aligned}$$

Let

$$z \equiv \frac{(y_{gi}-\mu_g)^2}{\sigma_g^2} - \frac{2\rho_{gh}(y_{gi}-\mu_g)(y_{hi}-\mu_h)}{\sigma_g\sigma_h} + \frac{(y_{hi}-\mu_h)^2}{\sigma_h^2}$$

Differentiating the function with respect to ρ_{gh}

$$\begin{aligned}&\nabla \log P(\rho_{gh}) \\ \propto &\nabla \log dMVN(y_{gi},y_{hi}) \\ \propto &\nabla \log \frac{1}{2\pi\sigma_g\sigma_h\sqrt{1-\rho_{gh}^2}} \exp\left[-\frac{z}{2\left(1-\rho_{gh}^2\right)}\right] \\ \propto &\nabla \left[-\frac{1}{2}\log(1-\rho_{gh}^2) - \frac{z}{2\left(1-\rho_{gh}^2\right)}\right] \\ \propto &\left[\frac{\rho_{gh}}{(1-\rho_{gh}^2)} + \frac{(y_{gi}-\mu_g)(y_{hi}-\mu_h)}{\sigma_g\sigma_h\left(1-\rho_{gh}^2\right)} - \frac{z\rho_{gh}}{\left(1-\rho_{gh}^2\right)^2}\right] \\ \propto &\left[\frac{\rho_{gh}}{\left(1-\rho_{gh}^2\right)} + \frac{(1-\rho_{gh})^2(y_{gi}-\mu_g)(y_{hi}-\mu_h)}{\sigma_g\sigma_h\left(1-\rho_{gh}^2\right)^2} + \frac{\rho_{gh}\left[\sigma_h(y_{gi}-\mu_g)-\sigma_g(y_{hi}-\mu_h)\right]^2}{\sigma_g^2\sigma_h^2\left(1-\rho_{gh}^2\right)^2}\right]\end{aligned}$$

If only one metabolite are fully observed or with less than 0.05 missing values

$$\begin{aligned}
\log P_{r_{gi}=1,r_{hi}=0,y_{gi}}(\rho_{gh}) &= \log P(r_{gi}=1, r_{hi}=0 \mid y_{gi}) P(y_{gi}) \\
&= \log \int P(r_{gi}=1, r_{hi}=0, y_{hi} \mid y_{gi})\, dy_{hi} \\
&= \log \int P(r_{gi}=1, r_{hi}=0 \mid y_{hi}, y_{gi}) P(y_{hi} \mid y_{gi})\, dy_{hi} \\
&\propto \log \int P(r_{hi}=0 \mid y_{hi}) P(y_{hi} \mid y_{gi})\, dy_{hi} \\
&\propto \log \left\{ \int \left(1 - \Psi\{\alpha_h(y_{hi} - \delta_h)\}\right) dMVN(y_{hi} \mid y_{gi})\, dy_{hi} \right\} \\
&\propto \log \left\{ \int \frac{1}{\sqrt{\pi}} \left(1 - \Psi\{z\}\right) e^{-\epsilon^2}\, d\epsilon \right\}
\end{aligned}$$

where

$$z = \alpha_h(\sqrt{2}\sigma_h \sqrt{1-\rho_{gh}^2}\epsilon + \mu_{hi} + \frac{\sigma_h}{\sigma_g}\rho_{gh}(y_{gi} - \mu_{gi}) - \delta_h)$$

Differentiating the function with respect to ρ_{gh}

$$\begin{aligned}
&\nabla_\rho \log P_{r_{gi}=1,r_{hi}=0,y_{gi}}(\rho_{gh}) \\
\propto &\nabla_\rho \log \left\{ \int \frac{1}{\sqrt{\pi}} \left(1 - \Psi\{z)\}\right) e^{-\epsilon^2}\, dy_{hi} \right\} \\
\propto &\frac{\int \frac{1}{\sqrt{\pi}} \nabla_\rho \left(1 - \Psi\{z)\}\right) e^{-\epsilon^2}\, d\epsilon}{\int \frac{1}{\sqrt{\pi}} \left(1 - \Psi\{z\}\right) e^{-\epsilon^2}\, d\epsilon} \\
\propto &\frac{\int \frac{1}{\sqrt{\pi}} \frac{z(\nu+1)}{\nu+z^2} \Psi\{z\} \alpha_h \left(-\sqrt{2}\sigma_h \frac{\rho}{\sqrt{1-\rho_{gh}^2}}\epsilon + \frac{\sigma_h}{\sigma_g}(y_{gi} - \mu_{gi}) \right) e^{-\epsilon^2}\, d\epsilon}{\int \frac{1}{\sqrt{\pi}} \left(1 - \Psi\{z\}\right) e^{-\epsilon^2}\, d\epsilon}
\end{aligned}$$

Both metabolite are with 0.05 - 0.5 missing values

$$\begin{aligned}
&\log P_{r_{gi}=r_{hi}=0}(\rho_{gh}) \\
= &\log \iint P(r_{gi}=0 \mid y_{gi}) P(r_{hi}=0 \mid y_{hi}) P(y_{gi}, y_{hi})\, dy_{gi} dy_{hi} \\
= &\log \iint \left(1 - \Psi\{\alpha_g(y_{gi} - \delta_g)\}\right) \left(1 - \Psi\{\alpha_h(y_{hi} - \delta_h)\}\right) dMVN(y_{gi}, y_{hi})\, dy_{gi} dy_{hi}
\end{aligned}$$

Differentiating the function with respect to ρ_{gh}

$$
\begin{aligned}
&\nabla_\rho \log P_{r_g=r_h=0}(\rho_{gh}) \\
\propto &\nabla_\rho \log \iint \left(1-\Psi\{\alpha_g(y_{gi}-\delta_g)\}\right)\left(1-\Psi\{\alpha_h(y_{hi}-\delta_h)\}\right) dMVN(y_{gi},y_{hi})\, dy_{gi}dy_{hi} \\
\propto &\frac{\nabla_\rho \iint \left(1-\Psi\{\alpha_g(y_{gi}-\delta_g)\}\right)\left(1-\Psi\{\alpha_h(y_{hi}-\delta_h)\}\right) dMVN(y_{gi},y_{hi})\, dy_{gi}dy_{hi}}{\iint \left(1-\Psi\{\alpha_g(y_{gi}-\delta_g)\}\right)\left(1-\Psi\{\alpha_h(y_{hi}-\delta_h)\}\right) dMVN(y_{gi},y_{hi})\, dy_{gi}dy_{hi}} \\
\propto &\frac{\iint \left(1-\Psi\{\alpha_g(y_{gi}-\delta_g)\}\right)\left(1-\Psi\{\alpha_h(y_{hi}-\delta_h)\}\right) dMVN(y_{gi},y_{hi}) g(\rho)\, dy_{gi}dy_{hi}}{\iint \left(1-\Psi\{\alpha_g(y_{gi}-\delta_g)\}\right)\left(1-\Psi\{\alpha_h(y_{hi}-\delta_h)\}\right) dMVN(y_{gi},y_{hi})\, dy_{gi}dy_{hi}}
\end{aligned}
$$

Bibliography

[1] Lin S Chen, Jiebiao Wang, Xianlong Wang, and Pei Wang. A mixed-effects model for incomplete data from labeling-based quantitative proteomics experiments. *The annals of applied statistics*, 11(1):114, 2017.

[2] Kieu Trinh Do, Simone Wahl, Johannes Raffler, Sophie Molnos, Michael Laimighofer, Jerzy Adamski, Karsten Suhre, Konstantin Strauch, Annette Peters, Christian Gieger, et al. Characterization of missing values in untargeted ms-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics*, 14(10):1–18, 2018.

[3] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[4] Chris McKennan, Carole Ober, Dan Nicolae, et al. Estimation and inference in metabolomics with nonrandom missing data and latent factors. *Annals of Applied Statistics*, 14(2):789–808, 2020.

[5] David B Richardson and Antonio Ciampi. Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *American journal of epidemiology*, 157(4):355–363, 2003.

[6] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.