Stony Brook University

Multi-Platform Comparison with Structural Equation Modeling and Errors-in-Variables Models with Random Loadings

A Dissertation Presented

by

Jinmiao Fu

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

May 2015

Stony Brook University

The Graduate School

Jinmiao Fu

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Wei Zhu – Dissertation Advisor
Deputy Chair, Professor, Department of Applied Mathematics and Statistics

Song Wu - Chairperson of Defense
Assistant Professor, Department of Applied Mathematics and Statistics

Xuefeng Wang
Assistant Professor, Department of Applied Mathematics and Statistics

Roman Kotov
Associate Professor, Department of Psychiatry

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Multi-Platform Comparison with Structural Equation Modeling and Errors-in-Variables Models with Random Loadings

by

Jinmiao Fu

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

2015

With the rapid advancement of biotechnology, multiple measurement platforms of microbiome abundance are increasingly available. These include the traditional platforms of gene microarray and quantitative PCR, as well as the modern next-generation sequencing technique. Consequently, the evaluation of the consistencies of these platforms has also become an increasingly crucial topic. Classic methods including using the Pearson correlation or the more suitable errors-in-variables (EIV) models to gauge the linear dependency between two platforms. Our group is among the leaders in applying the structural equation modeling (SEM) to estimate the relationships among three or more platforms and to combine these measurements for an optimal joint analysis. However, our previous work, as well as those of the others, only examines the agreement for each individual bacterium. In this thesis, we have developed a novel random coefficient SEM model to determine the agreement of different platforms across the entire microbiomes together taking into account the heterogeneity of individual bacterium.

We further applied this novel platform comparison method to a 16S ribosomal RNA sequencing study on bacteria abundance with three measurement modalities referred to as the V1V2, V1V3 and V3V4 windows. These are indeed three different targeting regions of primers when generating the amplicons. The newly developed SEM method with random loadings aims to test the average overall and pairwise consistency among these three platforms. Subsequently, good agreement between V1V2 and V3V4, and between V1V3 and V3V4 is found, while more discrepancy between V1V2 and V1V3 is detected. Moreover, the prediction of random loadings,

a by-product of the model above, is able to elucidate the performance of platforms on each individual bacterium.

The paradigm mentioned above could be easily adjusted to situations where only two platforms are available, which is another contribution of this work. Errors-in-variables (EIV) model with random coefficients (loadings) is proposed for the given task. To further confirm the conclusions above, pairwise comparison is performed and we are glad to report that coherent results are obtained.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor Prof. Wei Zhu sincerely, who not only offers me countless helps in my research, but also provides me with suggestions about career path in the future. Besides giving guidance on the topic in this dissertation, she also encourages me to browse various fields to prepare me for the current and the academic and industrial world. I am grateful to have her as my mentor, in research and in life.

My special thanks go to Prof. Roman Kotov and Prof. Evelyn Bromet for their supports in these years. It has always been a pleasure for me to work with them. My sincere thanks also go to Dr. Ellen Li for allowing me to use her data, and to Prof. Song Wu and Prof. Xuefeng Wang for taking time out of their busy research and teaching schedules to sit on my committee.

I also appreciate the helpful discussions and suggestions from my colleagues and friends Tian Feng and Dr. Yuanhao Zhang on my model. I also thank Erya Huang, Ruofeng Wen, Lu Zhao and all my academic siblings for their friendship and supports.

Last but not the least, I would like to thank my parents for raising me, supporting me, and loving me,unconditionally.

Chapter 1. Linear Mixed Model

1.1 Introduction

Linear mixed model (LME), also called multilevel model or random effect model, is a regression model suitable when repeated measures are made on the same unit longitudinally, or when units are divided into clusters. It could evaluate the overall linear relation between response and covariates, while allowing for heterogeneity within each unit or cluster of units. An important property of the LME is, unlike the simple or multiple linear regression where observations are assumed to be independent from each other, observations within the same cluster are correlated. Thanks to its ability to deal with missing data, it is always preferred over the repeated measure ANOVA.

1.2 Model Setting

Suppose the data contains N independent clusters, $i = 1, \cdots, N$, and each cluster has n_i measurements, $j = 1, \cdots, n_i$, with response variable Y_{ij}, covariates x_{ij} $(p \times 1)$ corresponding to fixed effects β, and z_{ij} $(q \times 1)$ corresponding to random effects b_i, and in most cases z_{ij} will be a subset of x_{ij}, then the model would be

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i \tag{1.2.1}$$

where $Y_i = (Y_{i1}, \cdots, Y_{in_i})$, $X_i = \left(x_{i1}, \cdots, x_{in_i}\right)^T$, $Z_i = \left(\mathrm{z_{i1}}, \cdots, \mathrm{z_{in_i}}\right)^{\mathrm{T}}$, $b_i \sim N(0, D)$ and $\varepsilon_i \sim N(0, R_i)$, with $R_i = \sigma^2 I_{n_i}$, usually.

To better understand the model above, suppose there are N schools in a certain area, and in the i^{th} school there are n_i students. Researchers are interested in studying the relation between each student's midterm score x_{ij} with his/her final score y_{ij}, where i is the school index, and j the student index. Since this relation will not be constant across all the schools, one can simple fit a simple linear regression by

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{bmatrix} = \begin{bmatrix} 1 & X_{i1} \\ \vdots & \vdots \\ 1 & X_{in_i} \end{bmatrix} \begin{bmatrix} \beta_{i0} \\ \beta_{i1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{in_i} \end{bmatrix} \tag{1.2.2}$$

for each school i separately, however this could be burdensome when N gets large, and more importantly, it is highly plausible that for some i, the corresponding sample size (e.g. total number of students) n_i could be small, which will make the regression in this school unreliable. As a result, to handle situations where N is large and certain n_i is small, one could assume $\beta_{i0} = \beta_0 + b_{i0}$ and $\beta_{i1} = \beta_1 + b_{i1}$, where $b_i = (b_{i0}, b_{i1})^T \sim N(0, D)$.

1.3 Estimation

Given the model settings above, it follows naturally that $Y_i \sim N(X_i\beta, V_i)$, where $V_i = Z_i D Z_i^T + R_i$, and the parameters to be estimated are $\theta = (\beta, D, R_i)$, so the log likelihood function would be

$$l = C - \frac{1}{2}\Sigma_{i=1}^{N}[log(|V_i|) + (Y_i - X_i\beta)^T V_i^{-1}(Y_i - X_i\beta)] \quad (1.3.1)$$

1.3.1 Maximum Likelihood Estimator (MLE)

From $\frac{\partial l}{\partial \beta} = 0$, it could be obtained that $\hat{\beta} = (\Sigma_{i=1}^{N} X_i^T V_i^{-1} X_i)^{-1} \cdot \Sigma_{i=1}^{N} X_i^T V_i^{-1} Y_i$, meaning that the MLE is completely determined by D and R_i, and various algorithms including Newton-Raphson [1] and Expectation Maximization (EM) [2] are available to solve for D and R.

1.3.2 Restricted Maximum Likelihood (REML)

To demonstrate the motivation of REML, suppose the data are represented by $X_1, \cdots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$, i.e. $X = (X_1, \cdots, X_n)^T \sim N(\mu 1_n, \sigma^2 I_n)$, where 1_n denotes the vector $(1, \cdots, 1)^T$ of length n and I_n denotes the n-dimensional identity matrix, then it is obvious that $\hat{\mu}_{MLE} = \bar{X}$, and $\hat{\sigma}^2_{MLE} = \frac{\Sigma_{i=1}^{n}(X_i - \hat{\mu}_{MLE})^2}{n}$. It is well known that $\hat{\sigma}^2_{MLE}$ is biased because estimating $\hat{\mu}_{MLE}$ will consume 1 degree of freedom, and hence $\frac{\Sigma_{i=1}^{n}(X_i - \hat{\mu}_{MLE})^2}{n-1}$ would be unbiased. However, if μ is known, then $\hat{\sigma}^2_{MLE} = \frac{\Sigma_{i=1}^{n}(X_i - \mu)^2}{n}$ would be unbiased, because there is no loss of degrees of freedom in 'estimating' μ.

In order to address the biasness problem when μ and σ^2 are both unknown, REML tries to find a matrix A of dimension $(n-1) \times n$ that maps X from R^n to R^{n-1}, and in the meanwhile to guarantee $A\mu = 0$, then it could be obtained that $Y \triangleq AX \sim N(0, \sigma^2 AA^T)$. As a result, since

the mean of Y, which is 0, becomes a known constant, the $\hat{\sigma}^2_{MLE}$ based on Y instead of X would be unbiased.

Patterson and Thompson (1971) proposed the formal procedure of applying REML in linear mixed model [3]. The basic idea is that, since $Y_i \sim N(X_i\beta, Z_iDZ_i^T + R_i)$, if there exists a matrix A, such that $AY_i \sim N(0, \Sigma_i)$, then the variance estimates would be unbiased, and then $\hat{\beta}$ could be obtained later. Since $R_i = \sigma^2 I_{n_i}$, then $V_i = Z_iDZ_i^T + R_i = \sigma^2\left(I_{n_i} + \frac{1}{\sigma^2}Z_iDZ_i^T\right) = \sigma^2 H_i$. In their work, they selected $S_i = I_{n_i} - X_i(X_i^TX_i)^{-1}X_i^T$ and $H_i = X_i^TH_i^{-1}$, because $E[S_iY_i] = 0$, i.e. $S_iX_i = 0$, and $cov(S_iY_i, Q_iY_i) = 0$, meaning that the log likelihood of Y_i, which is l_i, could be decomposed into the log likelihood of S_iY_i, i.e l_{i1}, and Q_iY_i, that is, l_{i2}.

The unbiased estimates of H_i could be obtained through maximizing $\Sigma_{i=1}^N l_{i1}$ because $E[S_iY_i] = 0$, and after obtaining $\widehat{H}_i$, estimates of β could be generated by maximizing $\Sigma_{i=1}^N l_{i2}$ assuming H_i is known, that is, $\hat{\beta} = \left(\Sigma_{i=1}^N X_i^T\widehat{H}_i^{-1}X_i\right)^{-1} \cdot \Sigma_{i=1}^N X_i^T\widehat{H}_i^{-1}Y_i$

1.3.3 Random Effect Prediction

After obtaining the estimations of all the parameters, i.e. $\hat{\beta}$, $\widehat{D}$ and $\widehat{R}_i$, all of the $b_i's$ can also be predicted if the linear relation within each cluster is of interest. It is worth noticing that $b_i's$ are random variables that are not included in the likelihood function, therefore they could not be predicted by the MLE or REML method.

Based on the normality assumption, it is not hard to see that $\begin{bmatrix} b_i \\ Y_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ X_i\hat{\beta} \end{bmatrix}, \begin{bmatrix} \widehat{D} & \widehat{D}Z_i^T \\ Z_i\widehat{D} & \widehat{V}_i \end{bmatrix}\right)$, from which it could be derived that

$$b_i|Y_i \sim N\left(\widehat{D}Z_i^T\widehat{V}_i^{-1}\left(Y_i - X_i\hat{\beta}\right), \widehat{D}^{-1} - \widehat{D}Z_i^T\widehat{V}_i^{-1}Z_i\widehat{D}\right) \quad (1.3.2)$$

As a result, the mean of this conditional distribution could be used as a prediction of b_i, i.e. $\widehat{D}Z_i^T\widehat{V}_i^{-1}\left(Y_i - X_i\hat{\beta}\right)$.

Chapter 2. Structural Equation Modeling

2.1 Introduction

Structural equation modeling (SEM) is a general analysis framework used to study the structure among variables including observed variables and latent variables, the latter defined as variables that could not be measured directly, for e.g., IQ, ability etc. Observed variables that are measurements of the latent variables are referred to as the indicators or manifest variables. Moreover, SEM could be viewed as a general modeling framework encompassing other methods such as regression, factor analysis, mixed model and errors in variables model etc.

2.2 Model Specification

A simple latent SEM model could be defined as follows, for each sample i:

$$X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{bmatrix} = \Gamma\xi_i + \varepsilon_i \tag{2.2.1}$$

where $\Lambda = \left(\lambda_1, \cdots, \lambda_p\right)^T$ is a $p \times 1$ coefficient vector, and $\varepsilon_i = \left(\varepsilon_{i1}, \cdots, \varepsilon_{ip}\right)^T$ is the $p \times 1$ residual vector. Figure 2.1 shows a diagram of the model above, where ξ_i, the circled variable, is the latent variable, while the rectangular variables, $X_{i1}, \cdots, X_{ip}$, are the observed variables, or manifest variables. Moreover, ξ_i, from which there are only arrows pointing out, is called exogenous variable, while all the others pointed to by arrows are called endogenous variables.

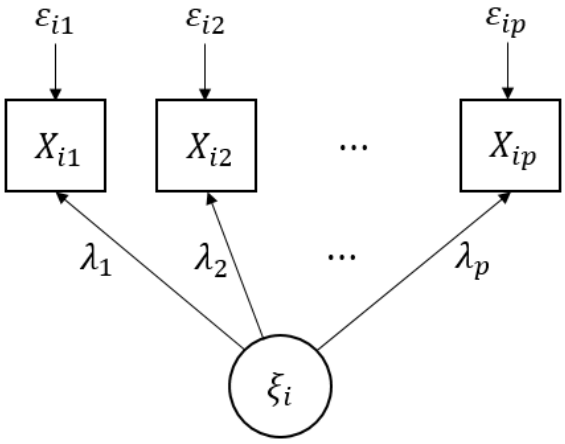

Figure 2.1. Diagram of model defined by Equation (2.2.1), where ξ_i is the latent factor and $X_{i1}, \cdots, X_{ip}$ are the corresponding manifest variables.

A convention of SEM is to center the data beforehand, i.e. substituting X_{ij} with $X_{ij} - \bar{X}_j$, where $\bar{X}_j = \frac{\Sigma_{i=1}^{N} X_{ij}}{N}$, which is why intercepts are absent from the model, and the mean of ξ_i will be zero. Given $\xi_i \sim N(0, \sigma_\xi^2)$, and $\varepsilon_{ij} \sim N(0, \sigma_j^2)$, the covariance matrix of X_i would be

$$V_i = \begin{bmatrix} \lambda_1^2\sigma_\xi^2 + \sigma_1^2 & \lambda_1\lambda_2\sigma_\xi^2 & \cdots & \lambda_1\lambda_p\sigma_\xi^2 \\ \lambda_1\lambda_2\sigma_\xi^2 & \lambda_2^2\sigma_\xi^2 + \sigma_2^2 & \cdots & \lambda_2\lambda_p\sigma_\xi^2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1\lambda_p\sigma_\xi^2 & \lambda_2\lambda_p\sigma_\xi^2 & \cdots & \lambda_p^2\sigma_\xi^2 + \sigma_p^2 \end{bmatrix} \tag{2.2.2}$$

from which it could be seen that there are infinite numbers of $\left(\{\lambda_i\}_{i=1,\cdots,p}, \{\sigma_i^2\}_{i=1,\cdots,p}, \sigma_\xi^2\right)$ that share the same V_i, because the scale, or unit, of the latent factor ξ_i could be arbitrary. As a result, for the purpose of model identification, two commonly used constraints are available, and they are (1) constrain λ_1 to be 1, and (2) constrain σ_ξ^2 to be 1.

Another identification issue will occur when there are only two manifest variables, i.e. $p = 2$. In this situation after applying a constraint above, e.g. $\lambda_1 = 1$, V_i will become $\begin{bmatrix} \sigma_\xi^2 + \sigma_1^2 & \lambda_2\sigma_\xi^2 \\ \lambda_2\sigma_\xi^2 & \lambda_2\sigma_\xi^2 + \sigma_2^2 \end{bmatrix}$, which contains four unknown parameters, λ_2, σ_ξ^2, σ_1^2 and σ_2^2, however, the valid information V_i provides is only three, thus the model is still non-identified. Consequently, another commonly used sufficient condition for the model to be identifiable is, each latent factor should have at least three manifest variables [4], i.e. $p \geq 3$.

SEM could handle more complicated model incorporating not only the relationship between manifest and latent variables, but also the relationship among latent variables or manifest variables themselves. A general setting of the SEM could be presented as the follows:

$$\eta_i = B\eta_i + \Gamma\xi_i + \zeta_i \tag{2.2.2}$$

$$Y_i = \Lambda_y\eta_i + \varepsilon_i \tag{2.2.3}$$

$$X_i = \Lambda_x\eta_i + \delta_i \tag{2.2.4}$$

and Figure 2.2 shows an example path diagram of the model above.

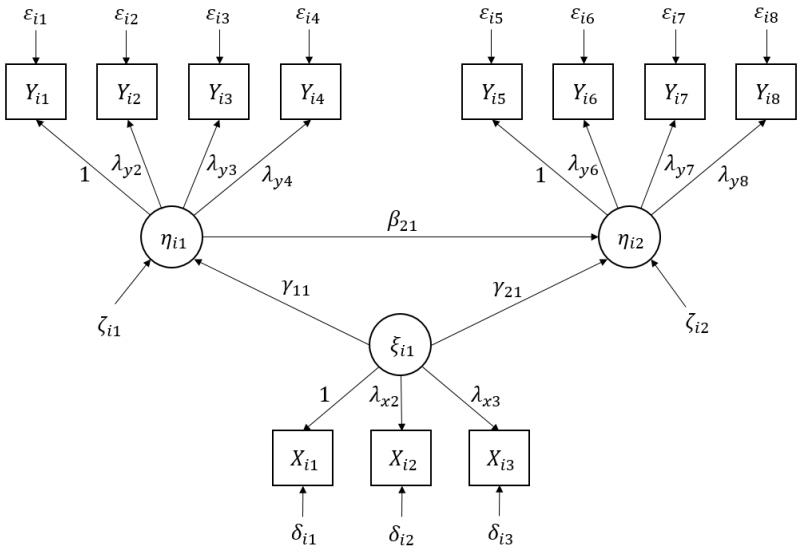

Figure 2.2. An example of SEM path diagram with complex structure including not only relationships between latent and manifest variables, but also relationships among latent variables themselves.

2.3 Estimation

From (2.2.2) – (2.2.4), it is easy to see that

$$\eta_i = (I - B)^{-1}\Gamma\xi_i + (I - B)^{-1}\zeta_i \tag{2.2.5}$$

$$y_i = (I - B)^{-1}\Gamma\xi_i + (I - B)^{-1}\zeta_i + \varepsilon_i \tag{2.2.6}$$

$$x_i = \xi_i + \delta_i \tag{2.2.7}$$

Subsequently, denoting the covariance matrix of ξ_i, ζ_i, δ_i and ε_i by V_ξ, V_ζ, V_δ and V_ε respectively, it follows naturally that

$$V_x \triangleq VAR(x) = V_\xi + V_\delta \tag{2.2.8}$$

$$V_y \triangleq VAR(y) = (I - B)^{-1}\left[\Gamma V_\xi \Gamma^T + V_\zeta\right]((I - B)^{-1})^T + V_\varepsilon \tag{2.2.9}$$

$$V_{xy} \triangleq COV(x, y) = V_{\xi}\Gamma^T((I - B)^{-1})^T \tag{2.2.10}$$

and the log likelihood of x and y would be

$$l \propto -\frac{1}{2}\Sigma_{i=1}^{N}\left(log|V| - \frac{1}{2}z_i^T V^{-1} z_i\right) \tag{2.2.11}$$

where $z_i = (x_i, y_i)^T$ and $V = \begin{bmatrix} V_x & V_{xy} \\ V_{xy}^T & V_y \end{bmatrix}$, and subsequently the MLE could be obtained.

Chapter 3. Expectation Maximization Algorithm

3.1 Introduction

The idea of Expectation-Maximization (EM) algorithm was established and named by Dempster, Laird and Rubin in 1977 [5]. Suppose Θ contains the parameters of interest, and $Y = (Y_1, \cdots, Y_n)$ are the observations, then the $\widehat{\Theta}_{MLE}$ that maximizes the log likelihood $l(\Theta|Y)$ could be cumbersome to solve. To overcome this, the EM algorithm assumes the existence of unobservable latent variables $X = (X_1, \cdots, X_n)$, which after being combined with Y would generate the completed observations $Z = (X, Y)$, and the corresponding log likelihood $l(\Theta|X, Y)$ has a neat form, then through the iteration between the E step and the M step, which would be covered later, solutions of $l(\Theta|Y)$ could be obtained upon convergence.

3.2 Procedure

The EM algorithm is achieved via the successive iteration between the Expectation Step (E step) and the Maximization Step (M step), where the E step is used to compute the conditional expectation of $l(\Theta|X, Y)$ given Y and $\hat{\Theta}^{(t)}$ at the current stage t, i.e. $E\left[l(\Theta|X, Y)|Y, \widehat{\Theta}^{(t)}\right]$, after which the M step will update $\hat{\Theta}^{(t)}$ by maximizing the conditional expectation obtained from the E step with respect to Θ, i.e. $\hat{\Theta}^{(t+1)} = \underset{\Theta}{argmax}\, E\left[l(\Theta|X, Y)|Y, \widehat{\Theta}^{(t)}\right]$. At the end, when the difference between two consecutive estimates, $\left|\hat{\Theta}^{(t+1)} - \hat{\Theta}^{(t)}\right|$, is less than a certain threshold Δ, usually $\Delta = 1e - 8$, the algorithm reaches convergence.

3.3 Application in linear mixed model

From (1.2.1), i.e. $Y_i = X_i\beta + Z_i b_i + \varepsilon_i$, $i = 1, \cdots, N$, $b_i \sim N(0, D)$ and $\varepsilon_i \sim N(0, \sigma^2 I_{n_i})$, in this case, the observations are $Y = (Y_1, \cdots, Y_N)$, and the latent variables are (b_i, ε_i), $i = 1, \cdots, N$, then the log likelihood of the complete observations is

$$l(\Theta|Y, b_i, \varepsilon_i) \propto -\frac{N}{2} log|D| - \frac{1}{2}\Sigma_{i=1}^{N} b_i^T D^{-1} b_i - \frac{1}{2}\Sigma_{i=1}^{N} log|R_i| - \frac{1}{2}\Sigma_{i=1}^{N} \varepsilon_i^T R_i^{-1} \varepsilon_i$$

$$= -\tfrac{N}{2} log|D| - \tfrac{1}{2} tr(D^{-1}\Sigma_{i=1}^{N} b_i b_i^T) - \tfrac{1}{2} M log\sigma^2 - \tfrac{1}{2\sigma^2}\Sigma_{i=1}^{N} \varepsilon_i^T \varepsilon_i \qquad (3.3.1)$$

where $M = \Sigma_{i=1}^{N} n_i$. If denoting $t_1 = \Sigma_{i=1}^{N} \varepsilon_i \varepsilon_i^T$ and $T_2 = \Sigma_{i=1}^{N} b_i b_i^T$, then based on the results from Davidian and Giltinan (1995) [6], it follows that

$$\tilde{t}_1^{(t)} \triangleq E\left[t_1|Y, \hat{\theta}^{(t)}\right] = \Sigma_{i=1}^{N}\left(\tilde{\varepsilon}_i^{(t)'}\tilde{\varepsilon}_i^{(t)} + tr\left(Cov\left\{\varepsilon_i|Y_i, \hat{\theta}^{(t)}\right\}\right)\right) \qquad (3.3.2)$$

$$\tilde{T}_2^{(t)} \triangleq E\left[T_2|Y, \hat{\theta}^{(t)}\right] = \Sigma_{i=1}^{N}\left(\tilde{b}_i^{(t)'}\tilde{b}_i^{(t)} + tr\left(Cov\left\{b_i|Y_i, \hat{\theta}^{(t)}\right\}\right)\right) \qquad (3.3.3)$$

where $\tilde{\varepsilon}_i^{(t)} \triangleq E\left[\varepsilon_i|Y_i, \hat{\theta}^{(t)}\right]$ and $\tilde{b}_i^{(t)} \triangleq E\left[b_i|Y_i, \hat{\theta}^{(t)}\right]$.

From (3.3.1) – (3.3.3), it is obvious that in the E step, we have

$$Q = E\left[l_c(\Theta|Y, b_i, \varepsilon_i)|Y, \hat{\theta}^{(t)}\right] = -\tfrac{N}{2} log|D| - \tfrac{1}{2} tr\left(D^{-1}\tilde{T}_2^{(t)}\right) - \tfrac{1}{2} M log\sigma^2 - \tfrac{1}{2\sigma^2}\tilde{t}_1^{(t)} \qquad (3.3.4)$$

Therefore in order to maximize Q in terms of D and σ^2, it could be obtained that

$$\hat{\sigma}^{2(t+1)} = \tfrac{\tilde{t}_1^{(t)}}{M} \text{ and } \hat{D}^{(t+1)} = \tfrac{\tilde{T}_2^{(t)}}{N}.$$

Chapter 4. Errors-in-Variables Model

4.1 Introduction

Errors-in-Variables (EIV) model, also called the measurement error model, is a regression model used to deal with situations where predictors/regressors are also subject to error. For example, when people are interested in the effects of fat intake during the last 24 hours have on certain response measure, the subjects may have to recall and estimate their fat intake, which is clearly error prone.

In classic regression model, for example, the simple linear regression, $Y = \beta_0 + \beta_1 X + \varepsilon$, estimated by the most popular ordinary least squares (OLS) method – if only the response Y is assumed to subject to error, while X is assumed be measured perfectly the OLS will fit β_0 and β_1 by minimizing the sum of squares of the vertical distances from each point to the regression line. Similarly, if X is the response, the sum of the squared horizontal distances will be minimized, as shown in Figure 4.1. The general EIV model, on the other hand, assuming errors exist in both the response and the regressor, would minimize the weighted sum of squared distances in both directions [7], and hence the entire class of EIV regression lines are always bounded by the two OLS regression lines of Y on X, and X on Y.

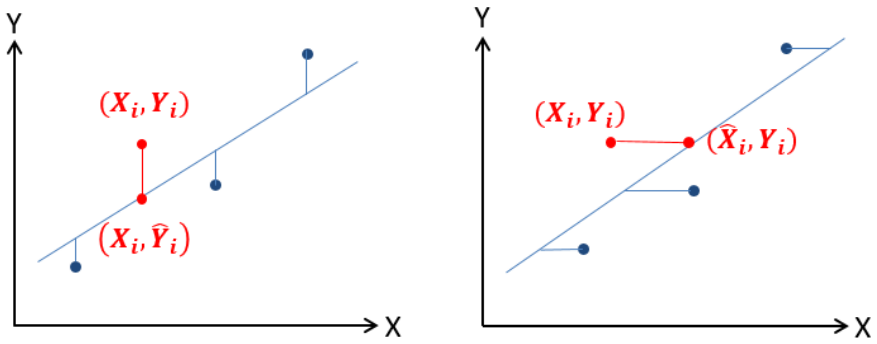

Figure 4.1. The ordinary least squares (OLS) regression line with Y as the error-prone response, and X as the error-free predictor (left); and similarly, the OLS with X as the response, and Y as predictor (right).

4.2 Model Setting

For each subject i, it is assumed that there exists the perfectly measured response η_i and predictor ξ_i, such that

$$\eta_i = \beta_0 + \beta_1 \xi_i \tag{4.2.1}$$

while the observed response and predictor satisfy

$$Y_i = \eta_i + \varepsilon_i \text{ and } X_i = \xi_i + \delta_i \tag{4.2.2}$$

where ε_i and δ_i are measurement errors with mean 0 and variances σ_ε^2 and σ_δ^2.

In terms of the true predictor ξ_i, distributional assumption on which may or may not be applied. If $\xi_i's$ are considered as fixed but unknown parameters, it is called functional relation, while if $\xi_i's$ are assumed to follow certain distribution, usually $\xi_i \sim N(\mu, \sigma_\xi^2)$, it becomes a structural one [8].

Sometimes the linear relationship between η_i and ξ_i indicated by (4.2.1) is not satisfied exactly, and thus there exists an equation error τ_i [9], meaning $\eta_i = \beta_0 + \beta_1\xi_i + \tau_i$, and this topic will be mentioned again in Section 5.6.

4.3 Estimation

4.3.1 Functional EIV

Since $\xi_i's$ are unknown parameters, the log likelihood would be

$$l \propto -\frac{N}{2}\left(log\sigma_\delta^2 + log\sigma_\varepsilon^2\right) - \frac{1}{2}\Sigma_{i=1}^{N}\left[\frac{(X_i-\xi_i)^2}{\sigma_\delta^2} + \frac{(Y_i-\beta_0-\beta_1\xi_i)^2}{\sigma_\varepsilon^2}\right] \tag{4.3.1}$$

where if all the $\xi_i's$ are constrained to be equal to $X_i's$, then all the corresponding term $\frac{(X_i-\xi_i)^2}{\sigma_\delta^2}$ will always be zero no matter how small σ_δ^2 is, thus as σ_δ^2 goes to zero, $-log\sigma_\delta^2$ will go to positive infinity, and so the whole log likelihood, therefore the model, is not identified.

4.3.2 Structural EIV

Since $\xi_i \sim N(\mu, \sigma_\xi^2)$, $Z_i = (X_i, Y_i)^T$ will follow bivariate normal with mean $\mu_Z = (\mu, \beta_0 + \beta_1\mu)^T$ and covariance matrix $V = \begin{bmatrix}\sigma_\xi^2 + \sigma_\delta^2 & \beta_1\sigma_\xi^2 \\ \beta_1\sigma_\xi^2 & \beta_1^2\sigma_\xi^2 + \sigma_\varepsilon^2\end{bmatrix}$, thus the log likelihood would be

$$l \propto -\frac{N}{2} log|V| - \frac{1}{2}\Sigma_{i=1}^{N}(Z_i - \mu_Z)^T V^{-1}(Z_i - \mu_Z) \tag{4.3.2}$$

Because V is a matrix containing three distinct elements but four unknown parameters, i.e. $(\beta_1, \sigma_\xi^2, \sigma_\delta^2, \sigma_\varepsilon^2)$, the model is also non-identifiable like functional EIV.

4.3.3 Identifiability

The pattern of EIV could be depicted by Figure 4.2, which is like the diagram of Figure 2.1, thus EIV could be considered as a special SEM with one latent factor and two manifest variables, which is clearly non-identifiable as explained in Section 2.2. For the purpose of identification, further constraint is needed, and the most commonly used one is to assume $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\delta^2}$ is known, then the MLE of parameters could be obtained as shown in Casella and Berger [9], where

$$\hat{\beta}_1 = \frac{S_{YY} - \lambda S_{XX} + \sqrt{(S_{YY} - S_{XX})^2 + 4\lambda S_{XY}^2}}{2S_{XY}} \text{ and } \hat{\beta}_0 = \bar{Y} - \hat{\beta}\bar{X} \tag{4.3.3}$$

with $S_{XX} = \Sigma_{i=1}^{N}(X_i - \bar{X})^2$, $S_{YY} = \Sigma_{i=1}^{N}(Y_i - \bar{Y})^2$ and $S_{XY} = \Sigma_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$. Furthermore, for functional EIV,

$$\hat{\xi}_i = \frac{\lambda X_i + \hat{\beta}_1(Y_i - \hat{\beta}_0)}{\lambda + \hat{\beta}_1^2} \tag{4.3.4}$$

$$\hat{\sigma}_\delta^2 = \frac{1}{2N(\lambda + \hat{\beta}_1^2)}\Sigma_{i=1}^{N}\left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right]^2 \text{ and } \hat{\sigma}_\varepsilon^2 = \lambda\hat{\sigma}_\delta^2 \tag{4.3.5}$$

while for structural EIV,

$$\hat{\mu} = \bar{X} \text{ and } \hat{\sigma}_{\xi}^2 = \frac{S_{XY}}{N\widehat{\beta}_1} \tag{4.3.6}$$

$$\hat{\sigma}_{\delta}^2 = \frac{1}{N}\left(S_{XX} - \frac{S_{XY}}{\widehat{\beta}_1}\right) \text{ and } \hat{\sigma}_{\varepsilon}^2 = \lambda\hat{\sigma}_{\delta}^2 \tag{4.3.7}$$

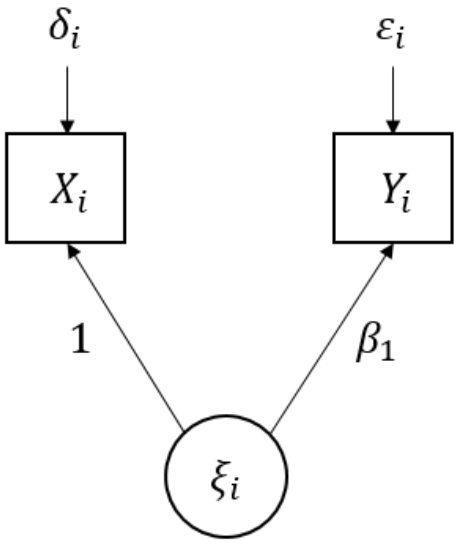

Figure 4.2. Diagram of a structural EIV model, which is equivalent to SEM with one latent factor and two corresponding manifest variables.

4.4 Choice of λ

Without preliminary knowledge it is hard for one to choose the correct λ, but there are two commonly used choices of λ that have wonderful geometric interpretations.

4.4.1 Orthogonal Regression

If $\lambda = 1$, then the MLE is identical to the model that tries to minimize the sum of the squared perpendicular distances between sample points and the fitted line as illustrated in Figure 4.3.

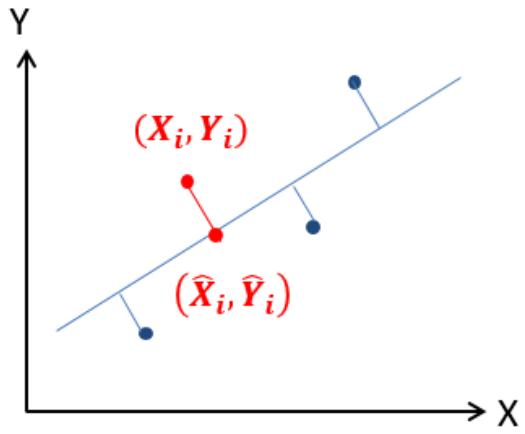

Figure 4.3. Fitting OR – minimizing the sum of squared perpendicular distances between the sample points and the fitted line

4.4.2 Geometric Mean Regression

If $\lambda = \frac{S_{YY}}{S_{XX}}$, then the MLE is identical to the model that minimizes the sum of the right triangular areas formulated by the sample points and the fitted line as illustrated in Figure 4.4.

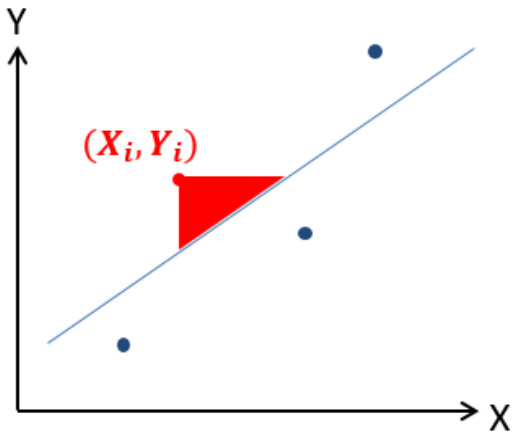

Figure 4.4. Fitting GMR – minimize the sum of areas of triangles formed by sample points and fitted line

4.5 Estimating λ

In reality it is often the case that one cannot determine which λ to use and has no available preliminary knowledge on λ, then with the help of replicates, i.e. each subject i is measured several times, the model could be identified without knowing λ.

4.5.1 Functional EIV

Barnett (1970) proposed the theoretical framework of functional EIV with replicates [10], where it was assumed that for each subject i, there are n_i replicates such that

$$\begin{cases} X_{ij} = \xi_i + \delta_{ij} \\ Y_{ij} = \eta_i + \varepsilon_{ij} \\ \eta_i = \beta_0 + \beta_1 \xi_i \end{cases} \tag{4.5.1}$$

where $i = 1, \cdots, N$ and $j = 1, \cdots, n_i$, then $M = \Sigma_{i=1}^{N} n_i$ will be the total number of observation. His work is more general than the distributional assumption proposed in Section 4.3 in a way that δ_{ij} and ε_{ij} are allowed to have different variances across i while their ratio is held constant, i.e. $\delta_{ij} \sim N(0, \sigma_i^2)$ and $\varepsilon_{ij} \sim N(0, \lambda\sigma_i^2)$. As a result, the log likelihood would be

$$l \propto -\Sigma_{i=1}^{N} n_i \, log\left(\sigma_i^2\sqrt{\lambda}\right) - \frac{1}{2}\Sigma_{i=1}^{N}\Sigma_{j=1}^{n_i}\left[\frac{\left(X_{ij}-\xi_i\right)^2}{\sigma_i^2} + \frac{\left(Y_{ij}-\beta_0-\beta_1\xi_i\right)^2}{\lambda\sigma_i^2}\right] \tag{4.5.2}$$

And then the vector of parameters to be estimated would be

$$\Theta = \left(\beta_0, \beta_1, \lambda, \{\xi_i, \sigma_i^2\}_{i=1,\cdots,N}\right)^T \tag{4.5.3}$$

After setting all of the related partial derivatives of l to zero, it is not hard to see that

$$\hat{\sigma}_i^2 = \frac{\Sigma_{j=1}^{n_i}\left[\left(X_{ij}-\hat{\xi}_i\right)^2 + \frac{\left(Y_{ij}-\hat{\beta}_0-\hat{\beta}_1\hat{\xi}_i\right)^2}{\hat{\lambda}}\right]}{2n_i} \tag{4.5.4}$$

$$\hat{\lambda} = \frac{\Sigma_{i=1}^{N}\Sigma_{j=1}^{n_i}\frac{\left(Y_{ij}-\hat{\beta}_0-\hat{\beta}_1\hat{\xi}_i\right)^2}{\hat{\sigma}_i^2}}{M} \tag{4.5.5}$$

$$\left(\bar{X}_i - \hat{\xi}_i\right) + \frac{\hat{\beta}_1\left(\bar{Y}_i-\hat{\beta}_0-\hat{\beta}_1\hat{\xi}_i\right)}{\hat{\lambda}} = 0 \tag{4.5.6}$$

$$\Sigma_{i=1}^{N}\frac{n_i\left(\bar{Y}_i-\hat{\beta}_0-\hat{\beta}_1\bar{X}_i\right)}{\hat{\lambda}\hat{\sigma}_i^2} \tag{4.5.7}$$

$$\Sigma_{i=1}^{N}\frac{n_i\hat{\xi}_i\left(\bar{Y}_i-\hat{\beta}_0-\hat{\beta}_1\bar{X}_i\right)}{\hat{\lambda}\hat{\sigma}_i^2} \tag{4.5.8}$$

from which $\hat{\xi}_i$ could be eliminated from the equation system and subsequently

$$\hat{\sigma}_i^2 = \frac{\Sigma_{j=1}^{n_i}\left[\left(X_{ij}-\bar{X}_i\right)^2+\frac{\left(Y_{ij}-\bar{Y}_i\right)^2}{\hat{\lambda}}\right]}{2n_i}+\frac{\left(\bar{Y}_i-\hat{\beta}_0-\hat{\beta}_1\bar{X}_i\right)^2}{2\hat{\lambda}\hat{\Delta}} \tag{4.5.9}$$

$$\hat{\lambda} = \frac{\Sigma_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_i\right)^2+\frac{n_i}{\hat{\Delta}^2}\left(\bar{Y}_i-\hat{\beta}_0-\hat{\beta}_1\bar{X}_i\right)^2}{M\Sigma_{i=1}^{N}\hat{\sigma}_i^{-2}} \tag{4.5.10}$$

where $\hat{\Delta} = 1+\frac{\hat{\beta}_1^2}{\hat{\lambda}}$, and if defining

$$\tilde{M} = \Sigma_{i=1}^{N}\frac{n_i}{\hat{\sigma}_i^2},\ \tilde{X}_i = \frac{n_i\bar{X}_i}{\hat{\sigma}_i^2} \text{ and } \tilde{Y}_i = \frac{n_i\bar{Y}_i}{\hat{\sigma}_i^2} \tag{4.5.11}$$

it could be obtained that

$$\hat{\beta}_1 = \frac{\tilde{S}_{YY}-\hat{\lambda}\tilde{S}_{XX}+\sqrt{\left(\tilde{S}_{YY}-\hat{\lambda}\tilde{S}_{XX}\right)^2+4\hat{\lambda}\tilde{S}_{XY}^2}}{2\tilde{S}_{XY}} \text{ and } \hat{\beta}_0 = \frac{\tilde{Y}-\hat{\beta}_1\tilde{X}}{\tilde{M}} \tag{4.5.12}$$

where

$$\tilde{X} = \Sigma_{i=1}^{N}\tilde{X}_i \text{ and } \tilde{Y} = \Sigma_{i=1}^{N}\tilde{Y}_i \tag{4.5.13}$$

$$\tilde{S}_{XX} = \Sigma_{i=1}^{N}\frac{\hat{\sigma}_i^2\tilde{X}_i^2}{n_i}-\frac{\tilde{X}^2}{\tilde{M}},\ \tilde{S}_{YY} = \Sigma_{i=1}^{N}\frac{\hat{\sigma}_i^2\tilde{Y}_i^2}{n_i}-\frac{\tilde{Y}^2}{\tilde{M}} \text{ and } \tilde{S}_{XX} = \Sigma_{i=1}^{N}\frac{\hat{\sigma}_i^2\tilde{X}_i\tilde{Y}_i}{n_i}-\frac{\tilde{X}\tilde{Y}}{\tilde{M}} \tag{4.5.14}$$

Then (4.5.9), (4.5.10) and (4.5.12) could be processed iteratively to generate the estimates $\left(\beta_0,\beta_1,\lambda,\{\sigma_i^2\}_{i=1,\cdots,N}\right)^T$.

It is worth noticing that in (4.5.12), $\hat{\beta}_1$ and $\hat{\beta}_0$ have the same structure as in (4.3.3), with the subtle differences that all of the N, X_i and Y_i involved in (4.3.3) have been reweighted by $\hat{\sigma}_i^2$ as in (4.5.11).

4.5.2 Structural EIV

In parallel to Barnett's work, Chan and Mak (1979) proposed the corresponding framework for structural EIV [11] with additional distributional assumption of ξ_i, i.e. $\xi_i \sim N(\mu, \sigma^2)$. In this work there are extra constraints, including $\sigma_1^2 = \cdots = \sigma_N^2 = \sigma_\delta^2$ and $n_1 = \cdots = n_N = r$.

Let $X_i = (X_{i1}, \cdots, X_{ir})^T$, $Y_i = (Y_{i1}, \cdots, Y_{ir})^T$, and $Z_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix}$ for $i = 1, \cdots, N$, then $Z_i \sim N(\mu_Z, V)$, where

$$\mu_Z = (\mu 1_r^T, (\beta_0 + \beta_1 \mu) 1_r^T)^T \text{ and } V = \begin{bmatrix} \sigma_\delta^2 I_r + \sigma^2 1_r 1_r^T & \beta_1 \sigma^2 1_r 1_r^T \\ \beta_1 \sigma^2 1_r 1_r^T & \lambda \sigma_\delta^2 1_r + \beta_1^2 \sigma^2 1_r 1_r^T \end{bmatrix} \quad (4.5.15)$$

then the log likelihood becomes

$$l \propto -\frac{1}{2} N log|V| - \frac{1}{2} \Sigma_{i=1}^N (Z_i - \mu_Z)^T V^{-1} (Z_i - \mu_Z) \quad (4.5.16)$$

It was proven in their work that if defining

$$\bar{X}_i = \frac{\Sigma_{j=1}^r X_{ij}}{r} \text{ and } \bar{Y}_i = \frac{\Sigma_{j=1}^r Y_{ij}}{r} \quad (4.5.17)$$

$$T_{XX} = \frac{\Sigma_{i=1}^N \Sigma_{j=1}^r X_{ij}^2}{Nr} \text{ and } T_{YY} = \frac{\Sigma_{i=1}^N \Sigma_{j=1}^r Y_{ij}^2}{Nr} \quad (4.5.18)$$

$$W_{XX} = \frac{\Sigma_{i=1}^N \Sigma_{j=1}^r (X_{ij} - \bar{X}_i)^2}{Nr} \text{ and } W_{YY} = \frac{\Sigma_{i=1}^N \Sigma_{j=1}^r (Y_{ij} - \bar{Y}_i)^2}{Nr} \quad (4.5.19)$$

$$S_{XX} = \frac{\Sigma_{i=1}^{N} \bar{X}_i^2}{N}, S_{YY} = \frac{\Sigma_{i=1}^{N} \bar{Y}_i^2}{N} \text{ and } S_{XY} = \frac{\Sigma_{i=1}^{N} \bar{X}_i \bar{Y}_i}{N} \quad (4.5.20)$$

then the MLE of β_1 is the root of equation $k_0\beta_1^4 + k_1\beta_1^3 + k_2\beta_1^2 + k_3\beta_1 + k_4 = 0$, where

$$k_0 = (r-1)S_{XX}S_{XY}T_{XX} \quad (4.5.21)$$

$$k_1 = rS_{XX}^2W_{YY} - (r-1)S_{XY}^2T_{XX} - (r-1)S_{XX}S_{YY}T_{XX} - rS_{XY}^2W_{XX} \quad (4.5.22)$$

$$k_2 = (3r-1)S_{XY}(S_{YY}W_{XX} - S_{XX}W_{YY}) \quad (4.5.23)$$

$$k_3 = rS_{XY}^2W_{YY} + (r-1)S_{XY}^2T_{YX} + (r-1)S_{XX}S_{YY}T_{YY} - rS_{YY}^2W_{XX} \quad (4.5.24)$$

$$k_4 = -(r-1)S_{XY}S_{YY}T_{YY} \quad (4.5.25)$$

if a real solution exists.

4.6 Application in platform comparison

In 2012, our team applied the EIV model to compare the consistency between qPCR and Microsphere (MS) [12] in terms of measuring gene expression level. The mainstream of perform platform comparison is via the Pearson correlation [13], which is a valid index measuring linear dependency, but it is not sophisticated enough to capture any bias. Since both platforms are obviously subject to measurement error, the EIV model seems to be a perfect fit [14].

4.6.1 Data Structure

Measurements of 18 pre-selected platelet related genes, including TGFB2, APP, LAPTM4B, HIST1H2AG, NGFRAP1, C20orf103, H3F3A, SRP72, ACOT7, WASF3, CLEC1B, RPL32, ACTB, CRYM, RPS20, HIST1H1A, TPM1, CTNS, from 50 subjects are available for both qPCR and MS, and each measurement has three technical replicates. Table 4.1A and B shows the structure of the data points for qPCR, and the ones for MS has the same pattern, where X_{ij}^{k} is the qPCR measurement for the i^{th} gene, j^{th} subject and k^{th} replicate for $i = 1, \cdots, 18$, $j = 1, \cdots, 50$ and $k = 1, \cdots, 3$. Then similarly Y_{ij}^{k} is the corresponding measurement from MS.

There is a caveat that before the comparison, measurements from both platforms should be transformed into the same units. To achieve that, two housekeeping genes, RPL32 and RPS20 were selected, and denote them as gene i_1 and i_2, then for each gene i, X_{ij}^{k} was transformed into $X_{ij}^{k} / \frac{\bar{X}_{i_1 j} + \bar{X}_{i_2 j}}{2}$, where $\bar{X}_{i_1 j} = \frac{X_{i_1 j}^{1} + X_{i_1 j}^{2} + X_{i_1 j}^{3}}{3}$ and $\bar{X}_{i_2 j} = \frac{X_{i_2 j}^{1} + X_{i_2 j}^{2} + X_{i_2 j}^{3}}{3}$. Similarly, Y_{ij}^{k} was transformed into $Y_{ij}^{k} / \frac{\bar{Y}_{i_1 j} + \bar{Y}_{i_2 j}}{2}$. An intuitive interpretation of this transformation is, instead of comparing the raw measurements for qPCR and MS, for each gene i and subject j, his or her measurements, divided by the mean of measurements of the two housekeeping genes from the same subject, were compared between two platforms.

Table 4.1A – Data structure of measurements of 18 genes and 50 subjects from qPCR.

qPCR (X)	Subject 1			…	Subject 50		
	R1	R2	R3		R1	R2	R3
TGFB2	X_{11}^{1}	X_{11}^{2}	X_{11}^{3}	…	$X_{1,50}^{1}$	$X_{1,50}^{2}$	$X_{1,50}^{3}$
⋮		⋮				⋮	
CTNS	$X_{18,1}^{1}$	$X_{18,1}^{2}$	$X_{18,1}^{3}$	…	$X_{18,50}^{1}$	$X_{18,50}^{2}$	$X_{18,50}^{3}$

Table 4.1B – Data structure of measurements of 18 genes and 50 subjects from MS.

MS (Y)	Subject 1			…	Subject 50		
	R1	R2	R3		R1	R2	R3
TGFB2	Y_{11}^1	Y_{11}^2	Y_{11}^3	…	$Y_{1,50}^1$	$Y_{1,50}^2$	$Y_{1,50}^3$
⋮		⋮				⋮	
CTNS	$Y_{18,1}^1$	$Y_{18,1}^2$	$Y_{18,1}^3$	…	$Y_{18,50}^1$	$Y_{18,50}^2$	$Y_{18,50}^3$

4.6.2 Analysis of all the genes

Suppose all of the measurements have been properly transformed, then in order to apply Barnett's method in Section 4.5.1, the sample mean of triplicates for each gene and each subjects was computed at first, i.e. $\bar{X}_{ij} \triangleq \frac{X_{ij}^1+X_{ij}^2+X_{ij}^3}{3}$ and $\bar{Y}_{ij} = \frac{Y_{ij}^1+Y_{ij}^2+Y_{ij}^3}{3}$, then it was assumed that $\bar{X}_{ij} = \xi_i + \delta_{ij}$, $\bar{Y}_{ij} = \eta_i + \varepsilon_{ij}$ and $\eta_i = \beta_0 + \beta_1\xi_i$ like in (4.5.1), where for each gene i, the 50 subjects were considered as 50 replicates. If it is further assumed that $\delta_{ij} \sim N(0, \sigma_i^2)$ and $\varepsilon_{ij} \sim N(0, \lambda\sigma_i^2)$, then Barnett's method could be applied exactly, and the conclusion of $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = 1$ would indicate the consistency between these two platforms.

Figure 4.5 is the scatterplot of all of the $(\bar{X}_{ij}, \bar{Y}_{ij})$ pairs with different symbols for different genes, from which it is clear that HIST1H1A is an outlier gene, so it was excluded from the following regression analysis where we compared the results of OLS with Y as response variables (OLS_Y), OLS with X as response variable (OLS_X), orthogonal regression (OR), geometric mean regression (GMR) and Barnett's method (Barnett_EIV).

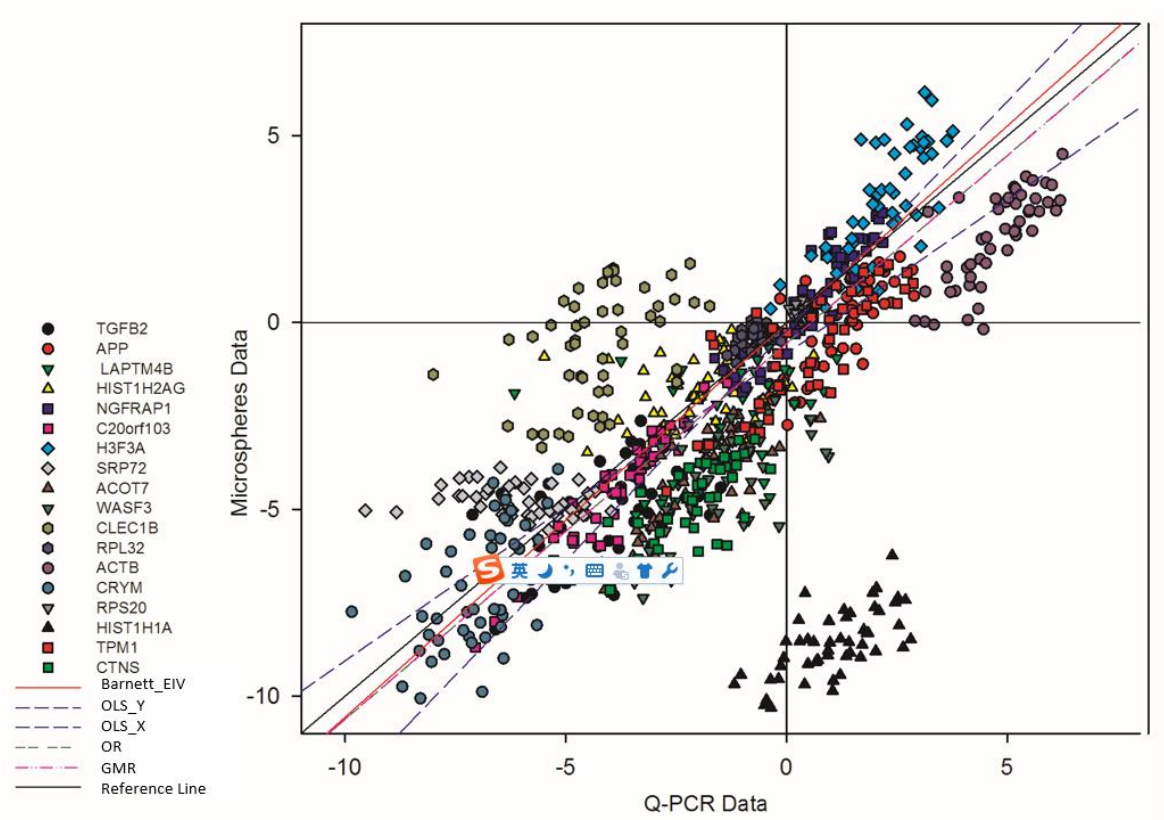

Figure 4.5. Scatter plots and Fitted lines from OLS_Y, OLS_X, OR, GMR, and Barnett_EIV of measurements from all the genes together.

Table 4.2 shows the results from all 5 methods with their corresponding λ, estimated $\hat{\beta}_0$, $\hat{\beta}_1$, and the bootstrapped confidence intervals [15]. Barnett's method is clearly superior in the sense that all the other four methods assume λ to be known, and its outputs, $\hat{\beta}_0 = -0.01$ and $\hat{\beta}_1 = 1.06$, strongly favors the conclusion that qPCR and MS are consistent.

Table 4.2. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of all the genes together.

	$\hat{\lambda}$	$\hat{\beta}_0$	$CI(\hat{\beta}_0)$	$\hat{\beta}_1$	$CI(\hat{\beta}_1)$
OLS_Y	∞	-0.83	(-0.95, -0.71)	0.82	(0.78, 0.86)
OLS_X	0	-0.23	(-0.42, -0.04)	1.23	(1.17, 1.29)
OR	1	-0.56	(-0.70, -0.43)	1.01	(0.95, 1.05)
GMR	1.01	-0.56	(-0.69, -0.44)	1.01	(0.97, 1.04)
Barnett_EIV	1.37E-06	-0.01	(-0.68, 1.20)	1.06	(0.48, 2.35)

4.6.3 Analysis of each individual gene

Although these two platforms are consistent in terms of all the genes together, it is obvious that for some genes, based on Figure 4.2, their scatter plots are far away from the reference line, i.e. $Y = X$, which makes it doubtful that whether assuming all the genes have the same linear pattern is plausible or not. As a result, EIV analysis on each individual gene could be done in a similar manner.

For each gene i, since replicates are available for each subject, we applied method used to Linnet (1993) to estimate λ at first [16], which is $\hat{\lambda}_i = \frac{\Sigma_{j=1}^{50}\Sigma_{k=1}^{3}\left(Y_{ij}^{k}-\bar{Y}_{ij}\right)^2}{\Sigma_{j=1}^{50}\Sigma_{k=1}^{3}\left(X_{ij}^{k}-\bar{X}_{ij}\right)^2}$, and used this $\hat{\lambda}_i$ to perform EIV on $\left\{\left(\bar{X}_{ij},\bar{Y}_{ij}\right)\right\}_{j=1,\cdots,50}$, hereafter abbreviated as Best_EIV. Similarly, OLS_Y, OLS_X, OR, GMR were also adopted on the same data points. Table 4.3A–Q shows the results for each gene in the same way Table 4.1 does, and Figure 4.6A–Q are the corresponding plots with the sample correlation of $\left\{\left(\bar{X}_{ij},\bar{Y}_{ij}\right)\right\}_{j=1,\cdots,50}$ at the top.

Table 4.3A. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of C20orf103.

C20orf103	$\hat{\lambda}$	$\hat{\beta}_0$	$\mathrm{CI}(\hat{\beta}_0)$	$\hat{\beta}_1$	$\mathrm{CI}(\hat{\beta}_1)$
OLS_Y	∞	0.19	(-0.37, 0.75)	1.26	(1.12, 1.39)
OLS_X	0	0.92	(-0.34, -0.04)	1.43	(1.29, 1.61)
OR	1	0.65	(0.04, 1.26)	1.37	(1.22, 1.52)
GMR	1.80	0.54	(-0.06, 1.14)	1.34	(1.20, 1.49)
Best_EIV	0.76	0.70	(0.08, 1.31)	1.38	(1.23, 1.53)

Table 4.3B. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of NGFRAP1.

NGFRAP1	$\hat{\lambda}$	$\hat{\beta}_0$	CI($\hat{\beta}_0$)	$\hat{\beta}_1$	CI($\hat{\beta}_1$)
OLS_Y	∞	-0.05	(-0.16, 0.07)	1.19	(1.03, 1.34)
OLS_X	0	-0.22	(-0.37, -0.10)	1.43	(1.27, 1.65)
OR	1	-0.15	(-0.28, -0.03)	1.34	(1.16, 1.51)
GMR	1.70	-0.13	(-0.25, -0.01)	1.30	(1.13, 1.48)
Best_EIV	0.48	-0.18	(-0.31, -0.05)	1.38	(1.19, 1.56)

Table 4.3C. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of TPM1.

TPM1	$\hat{\lambda}$	$\hat{\beta}_0$	CI($\hat{\beta}_0$)	$\hat{\beta}_1$	CI($\hat{\beta}_1$)
OLS_Y	∞	-1.13	(-1.22, -1.04)	0.91	(0.71, 1.11)
OLS_X	0	-1.36	(-1.45, -1.21)	1.43	(1.09, 1.61)
OR	1	-1.25	(-1.34, -1.17)	1.18	(0.98, 1.38)
GMR	1.30	-1.24	(-1.30, -1.17)	1.14	(1.00, 1.28)
Best_EIV	0.77	-1.27	(-1.35, -1.17)	1.22	(0.99, 1.40)

Table 4.3D. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of ACTB.

ACTB	$\hat{\lambda}$	$\hat{\beta}_0$	CI($\hat{\beta}_0$)	$\hat{\beta}_1$	CI($\hat{\beta}_1$)
OLS_Y	∞	-2.90	(-4.14, -1.66)	1.09	(0.82, 1.35)
OLS_X	0	-6.45	(-8.14, -3.81)	1.84	(1.28, 2.19)
OR	1	-5.13	(-6.49, -3.02)	1.56	(1.11, 1.84)
GMR	1.99	-4.44	(-5.37, -3.27)	1.41	(1.16, 1.61)
Best_EIV	1.41	-4.80	(-6.21, -2.97)	1.49	(1.10, 1.79)

Table 4.3E. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of ACOT7.

ACOT7	$\hat{\lambda}$	$\hat{\beta}_0$	CI($\hat{\beta}_0$)	$\hat{\beta}_1$	CI($\hat{\beta}_1$)
OLS_Y	∞	-2.30	(2.70, -1.90)	0.93	(0.73, 1.13)
OLS_X	0	-1.28	(-1.91, -0.85)	1.44	(1.12, 1.66)
OR	1	-1.76	(-2.25, -1.40)	1.20	(0.95, 1.38)
GMR	1.34	-1.85	(-2.35, -1.35)	1.16	(0.91, 1.41)
Best_EIV	4.02	-2.11	(-2.55, -1.67)	1.02	(0.80, 1.25)

Table 4.3F. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of APP.

APP	$\hat{\lambda}$	$\hat{\beta}_0$	$CI(\hat{\beta}_0)$	$\hat{\beta}_1$	$CI(\hat{\beta}_1)$
OLS_Y	∞	-1.34	(-1.68, -1.00)	1.06	(0.80, 1.33)
OLS_X	0	-2.35	(-3.13, -1.88)	1.86	(1.49, 2.48)
OR	1	-1.97	(-2.46, -1.47)	1.56	(1.17, 1.95)
GMR	1.98	-1.77	(-2.22, -1.33)	1.41	(1.05, 1.76)
Best_EIV	0.83	-2.02	(-2.35, -1.51)	1.60	(1.20, 2.00)

Table 4.3G. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of CTNS.

CTNS	$\hat{\lambda}$	$\hat{\beta}_0$	$CI(\hat{\beta}_0)$	$\hat{\beta}_1$	$CI(\hat{\beta}_1)$
OLS_Y	∞	-2.84	(-3.33, -2.35)	0.96	(0.74, 1.17)
OLS_X	0	-1.54	(-2.23, -1.08)	1.53	(1.23, 1.73)
OR	1	-2.13	(-2.66, -1.70)	1.27	(1.04, 1.46)
GMR	1.47	-2.27	(-2.63, -1.95)	1.21	(1.05, 1.35)
Best_EIV	0.13	-1.65	(-2.40, -1.21)	1.48	(1.15, 1.68)

Table 4.3H. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of H3F3A.

H3F3A	$\hat{\lambda}$	$\hat{\beta}_0$	$CI(\hat{\beta}_0)$	$\hat{\beta}_1$	$CI(\hat{\beta}_1)$
OLS_Y	∞	0.51	(-0.16, -1.19)	1.25	(0.94, 1.56)
OLS_X	0	-1.48	(-3.05, -0.55)	2.17	(1.74, 2.90)
OR	1	-0.89	(-1.92, 0.13)	1.90	(1.43, 2.37)
GMR	2.72	-0.35	(-1.24, 0.54)	1.65	(1.24, 2.06)
Best_EIV	0.04	-1.45	(-2.62, -0.29)	2.16	(1.62, 2.70)

Table 4.3I. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of TGFB2.

TGFB2	$\hat{\lambda}$	$\hat{\beta}_0$	$CI(\hat{\beta}_0)$	$\hat{\beta}_1$	$CI(\hat{\beta}_1)$
OLS_Y	∞	-2.53	(-3.54, -1.63)	0.61	(0.35, 0.84)
OLS_X	0	1.62	(-0.19, 5.69)	1.68	(1.22, 2.74)
OR	1	-0.94	(-2.46, 0.59)	1.02	(0.63, 1.42)
GMR	1.03	-0.97	(-2.48, 0.54)	1.01	(0.63, 1.40)
Best_EIV	1.88	-1.64	(-2.89, -0.39)	0.84	(0.52, 1.16)

Table 4.3J. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of WASF3.

WASF3	$\hat{\lambda}$	$\hat{\beta}_0$	CI($\hat{\beta}_0$)	$\hat{\beta}_1$	CI($\hat{\beta}_1$)
OLS_Y	∞	-3.04	(-3.44, -2.65)	0.82	(0.56, 1.08)
OLS_X	0	-1.57	(-2.70, -1.04)	1.79	(1.04, 2.15)
OR	1	-2.27	(-2.91, -1.64)	1.33	(0.91, 1.75)
GMR	1.47	-2.45	(-3.03, 0.54)	1.21	(0.83, 1.60)
Best_EIV	0.33	-1.86	(-2.68, -1.37)	1.60	(1.06, 1.92)

Table 4.3K. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of CRYM.

CRYM	$\hat{\lambda}$	$\hat{\beta}_0$	CI($\hat{\beta}_0$)	$\hat{\beta}_1$	CI($\hat{\beta}_1$)
OLS_Y	∞	-1.68	(-4.33, 0.98)	0.78	(0.40, 1.16)
OLS_X	0	13.78	(6.91, 33.87)	3.01	(2.02, 5.90)
OR	1	8.06	(0.63, 15.49)	2.18	(1.11, 3.25)
GMR	2.35	3.54	(-1.67, 8.75)	1.53	(0.78, 2.28)
Best_EIV	1.77	5.05	(-0.90,11.01)	1.75	(0.89, 2.61)

Table 4.3L. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of RPL32.

RPL32	$\hat{\lambda}$	$\hat{\beta}_0$	CI($\hat{\beta}_0$)	$\hat{\beta}_1$	CI($\hat{\beta}_1$)
OLS_Y	∞	-0.09	(-0.24, 0.06)	0.41	(0.18, 0.64)
OLS_X	0	0.91	(0.46, 2.52)	1.93	(1.24, 4.38)
OR	1	0.15	(-0.13, 0.44)	0.77	(0.34, 1.21)
GMR	0.79	0.23	(-0.10, 0.55)	0.89	(0.39, 1.39)
Best_EIV	0.32	0.55	(0.04, 1.06)	1.38	(0.61, 2.15)

Table 4.3M. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of LAPTM4B.

LAPTM4B	$\hat{\lambda}$	$\hat{\beta}_0$	CI($\hat{\beta}_0$)	$\hat{\beta}_1$	CI($\hat{\beta}_1$)
OLS_Y	∞	-2.17	(-2.57, -1.77)	0.44	(0.17, 0.72)
OLS_X	0	0.71	(-5.68, 1.88)	2.41	(-1.94, 3.21)
OR	1	-1.23	(-2.29, -0.15)	1.08	(0.36, 1.82)
GMR	1.07	-1.30	(-1.76, -0.95)	1.03	(0.72, 1.28)
Best_EIV	0.47	-0.42	(-2.14, 0.33)	1.64	(0.46, 2.15)

Table 4.3N. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of CLEC1B.

CLEC1B	$\hat{\lambda}$	$\hat{\beta}_0$	$CI(\hat{\beta}_0)$	$\hat{\beta}_1$	$CI(\hat{\beta}_1)$
OLS_Y	∞	1.63	(0.15, 2.71)	0.54	(0.20, 0.80)
OLS_X	0	10.91	(6.65, 26.98)	2.72	(1.72, 6.49)
OR	1	5.86	(2.06, 9.66)	1.54	(0.64, 2.43)
GMR	1.48	4.50	(1.49, 7.51)	1.22	(0.51, 1.92)
Best_EIV	0.13	10.10	(3.84, 16.36)	2.53	(1.06, 4.00)

Table 4.3O. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of SRP72.

SRP72	$\hat{\lambda}$	$\hat{\beta}_0$	$CI(\hat{\beta}_0)$	$\hat{\beta}_1$	$CI(\hat{\beta}_1)$
OLS_Y	∞	-5.27	(-5.81, -4.72)	-0.08	(-0.17, 0.01)
OLS_X	0	-12.50	(-32.24,19.67)	-1.28	(-4.56, 4.05)
OR	1	-5.32	(-5.93, -4.71)	-0.09	(-0.19, 0.01)
GMR	0.10	-6.73	(-8.89, -4.56)	-0.32	(-0.68, 0.04)
Best_EIV	0.08	-7.54	(-10.44, -1.54)	-0.46	(-0.94, 0.54)

Table 4.3P. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of HIST1H2AG.

HIST1H2AG	$\hat{\lambda}$	$\hat{\beta}_0$	$CI(\hat{\beta}_0)$	$\hat{\beta}_1$	$CI(\hat{\beta}_1)$
OLS_Y	∞	-1.29	(-1.66, -0.92)	0.22	(0.03, 0.41)
OLS_X	0	2.52	(-17.89,4.71)	2.22	(-8.49, 3.37)
OR	1	-1.00	(-1.62, -0.38)	0.37	(0.05, 0.70)
GMR	0.49	-0.38	(-1.54, 0.79)	0.70	(0.09, 1.31)
Best_EIV	0.19	1.10	(-8.50, 2.32)	1.48	(-3.56, 2.1)

Table 4.3Q. Estimates of λ, β_0, β_1 and their confidence intervals by OLS_Y, OLS_X, OR, GMR and Best_EIV on measurements of RPS20.

RPS20	$\hat{\lambda}$	$\hat{\beta}_0$	$CI(\hat{\beta}_0)$	$\hat{\beta}_1$	$CI(\hat{\beta}_1)$
OLS_Y	∞	0.19	(0.05, 0.32)	0.19	(-0.16, 0.55)
OLS_X	0	-2.76	(-0.80, 3.84)	7.94	(-9.40, 2.79)
OR	1	-0.90	(-3.05, 1.24)	3.06	(-2.58, 8.69)
GMR	1.52	-0.21	(-1.07, 0.66)	1.23	(-1.04, 3.51)
Best_EIV	0.07	-2.62	(-7.94, 2.70)	7.57	(-6.39,21.53)

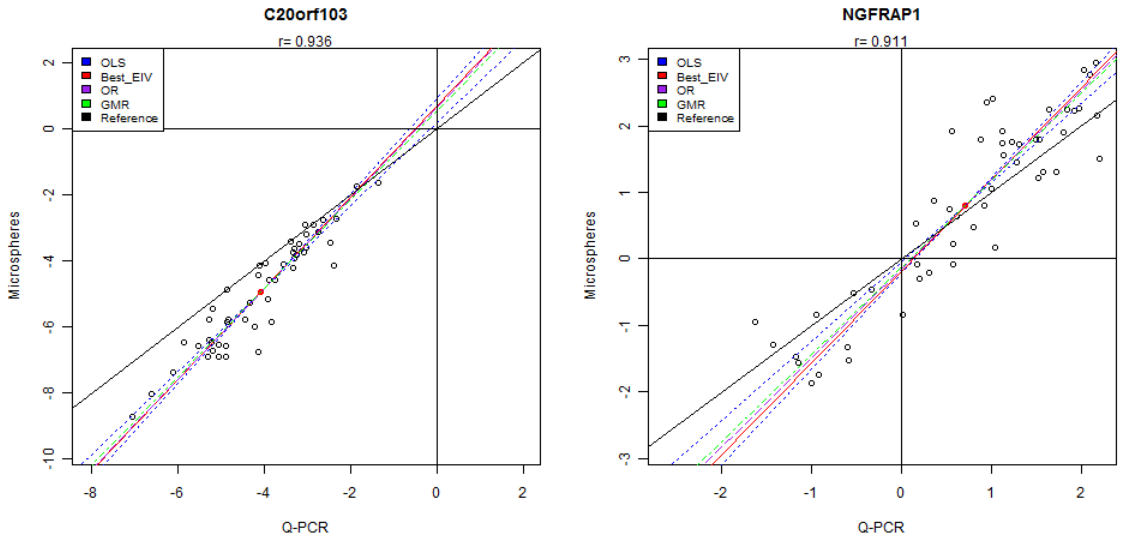

Figure 4.6. A (left) – Fitted lines from OLS_Y, OLS_X, OR, GMR, and Barnett_EIV of measurements from C20orf103; B (right) – corresponding plot from NGFRAP1.

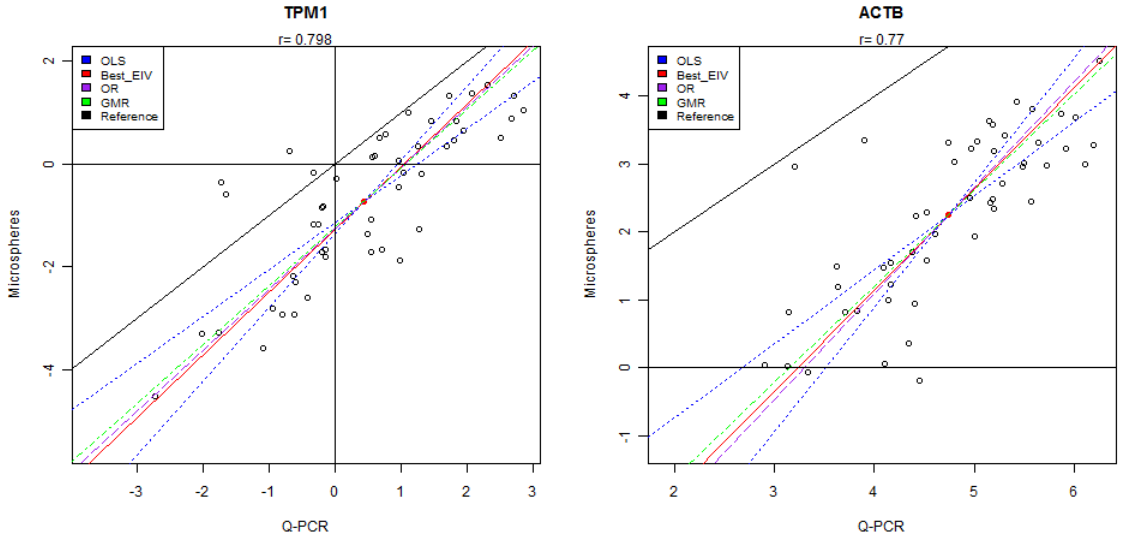

Figure 4.6. C (left) – Fitted lines from OLS_Y, OLS_X, OR, GMR, and Barnett_EIV of measurements from TPM1; D (right) – corresponding plot from ACTB.

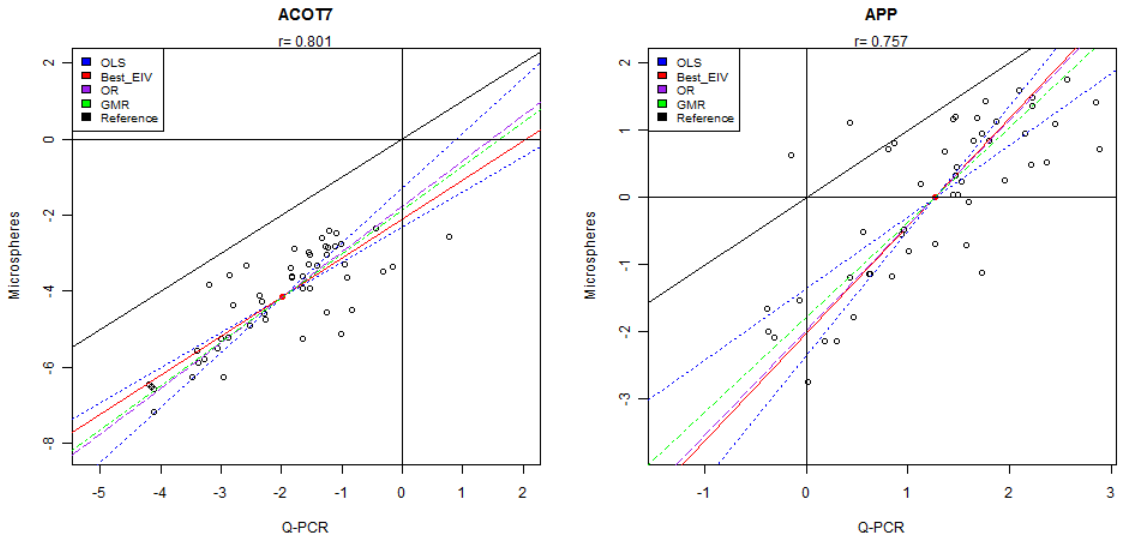

Figure 4.6. E (left) – Fitted lines from OLS_Y, OLS_X, OR, GMR, and Barnett_EIV of measurements from ACOT7; F (right) – corresponding plot from APP.

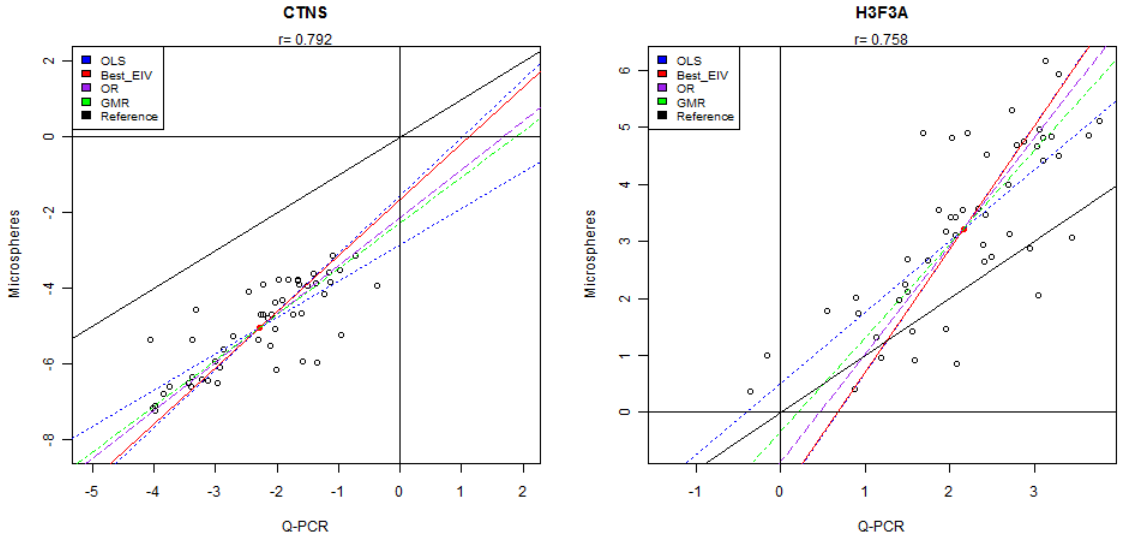

Figure 4.6. G (left) – Fitted lines from OLS_Y, OLS_X, OR, GMR, and Barnett_EIV of measurements from CTNS; H (right) – corresponding plot from H3F3A.

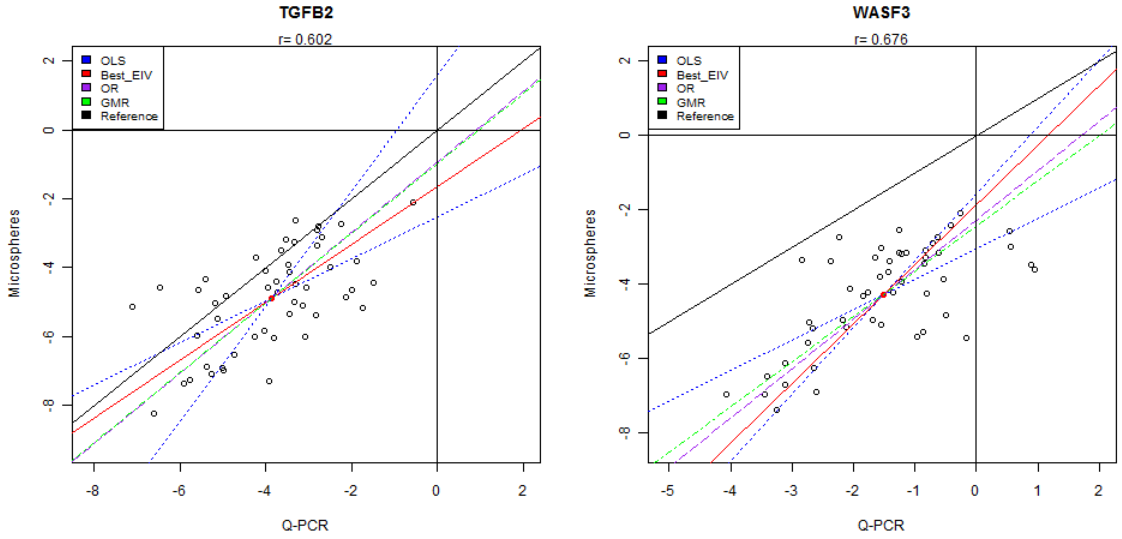

Figure 4.6. I (left) – Fitted lines from OLS_Y, OLS_X, OR, GMR, and Barnett_EIV of measurements from TGFB2; J (right) – corresponding plot from WASF3.

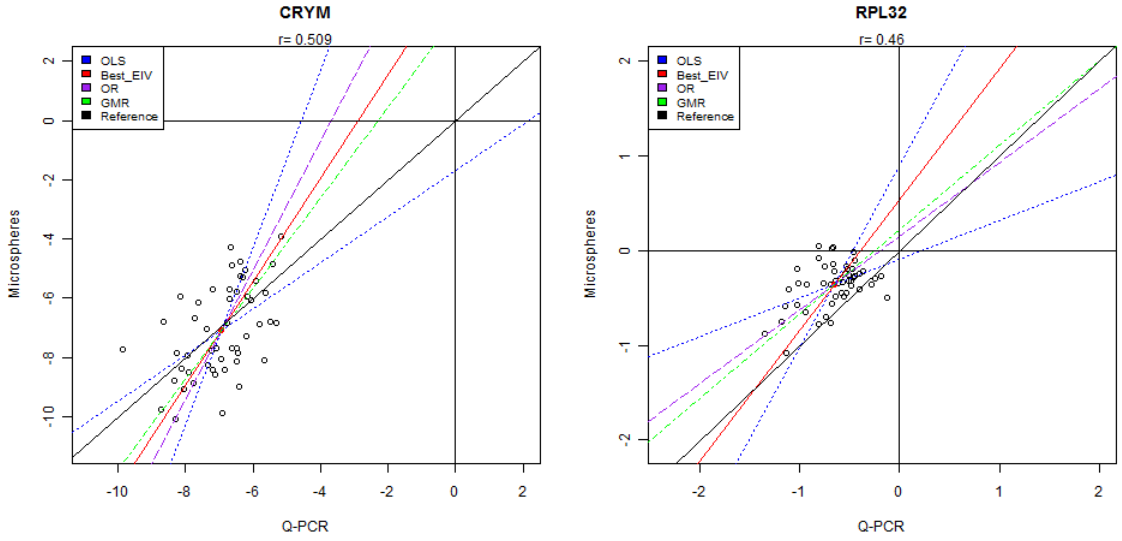

Figure 4.6. K (left) – Fitted lines from OLS_Y, OLS_X, OR, GMR, and Barnett_EIV of measurements from CRYM; L (right) – corresponding plot from RPL32.

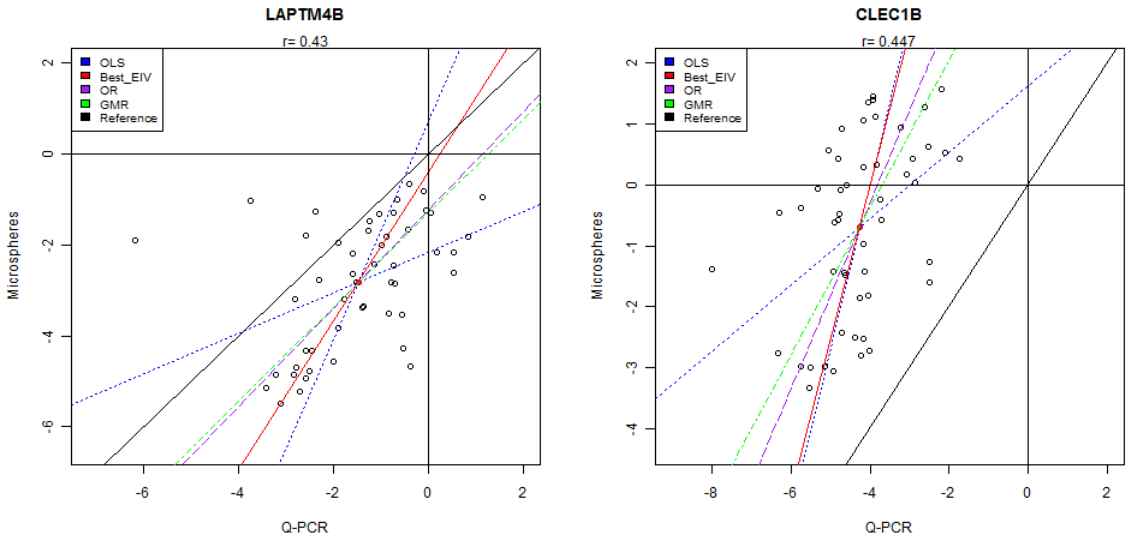

Figure 4.6. M (left) – Fitted lines from OLS_Y, OLS_X, OR, GMR, and Barnett_EIV of measurements from LAPTM4B; N (right) – corresponding plot from CLEC1B.

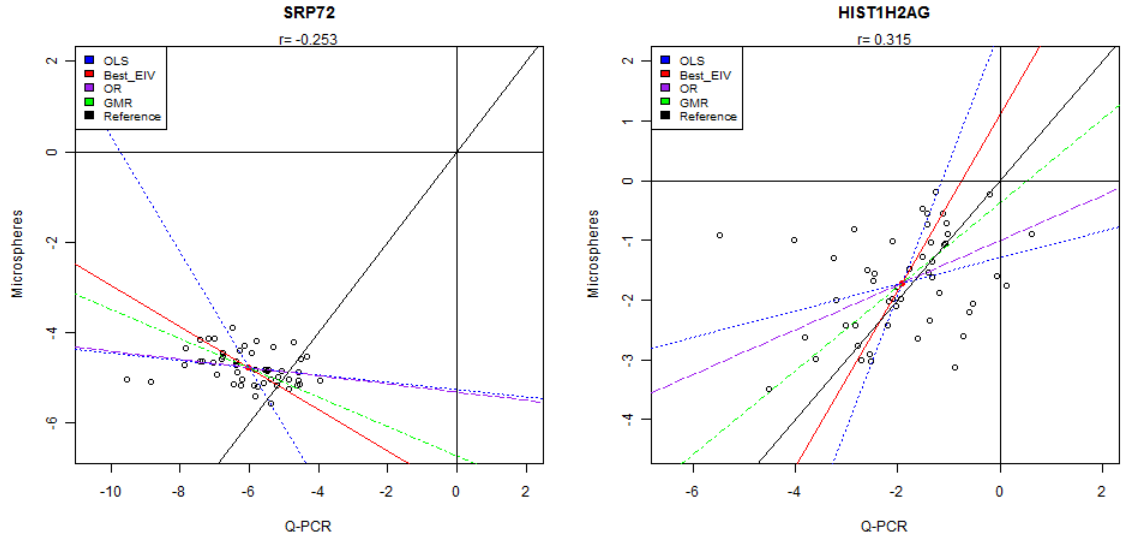

Figure 4.6. O (left) – Fitted lines from OLS_Y, OLS_X, OR, GMR, and Barnett_EIV of measurements from SRP72; P (right) – corresponding plot from HIST1H2AG.

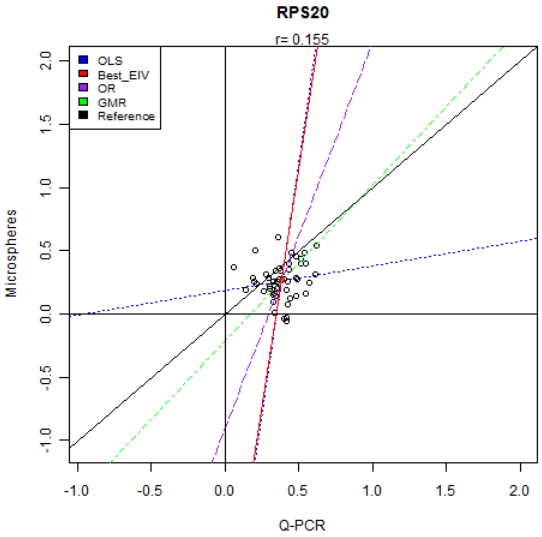

Figure 4.6Q. Fitted lines from OLS_Y, OLS_X, OR, GMR, and Barnett_EIV of measurements from RPS20.

4.6.4 Discussion

The outputs above indicate that the choice of λ will affect the judgment to a very large extent, and most genes are not consistent between qPCR and MS when analyzed separately, e.g. H3F3A, $CI(\hat{\beta}_0) = (-2.62, -0.29)$, which does not include 0, and $CI(\hat{\beta}_1) = (1.62, 2.70)$, which also does not include 1, thus it would be assertive to assume all the genes have the same pattern like in Section 4.6.2. Hence, a method that could analyze the whole gene cohort while allowing each gene to have individual pattern is needed, and the corresponding details will be covered in Section 7.7.

Chapter 5. Generalized Method of Moments

5.1 Introduction

The material of this chapter is enlightened by and organized as the same structure as the lecture note by University of Washington [17]. Generalized method of moments (GMM) was formulized by Hansen LP (1982) [18], which unlike MLE, does not rely on the knowledge of the distribution of data, and is widely used in finance and econometrics, besides it is usually computationally easy. The estimation is achieve via the orthogonal conditions from instrumental variables and residuals that would be described later, and the results have good properties like consistency and asymptotic efficiency.

5.2 Orthogonal Conditions

Given a linear regression model

$$y_i = x_i^T \beta + \varepsilon_i, i = 1, \cdots, N \tag{5.2.1}$$

where x_i is the L dimensional explanatory vector of the i^{th} observation, $\beta = (\beta_0, \cdots, \beta_{L-1})$ is the L dimensional coefficients of interest, and ε_i is the corresponding residual. The GMM assumes the existence of a K instrumental variables z_i for each i, which contain some or all of the elements in x_i are uncorrelated with the residual, i.e. $E[z_i \varepsilon_i] = 0$. Since z_i is K dimensional, $E[z_i \varepsilon_i] = 0$ is referred to as the K orthogonal conditions.

5.3 Estimation

Denoting w_i is the vector of unique elements in $\{y_i, x_i, z_i\}$, then from (4.2.1), the orthogonal conditions could be expressed as

$$E[g_i(w_i, \beta)] = E[z_i \varepsilon_i] = E[z_i(y_i - x_i^T \beta)] = 0 \tag{5.3.1}$$

where $g_i(w_i, \beta) \triangleq z_i \varepsilon_i = z_i(y_i - x_i^T \beta)$.

Expanding (5.3.1) will give us

$$\Sigma_{zy} = \Sigma_{zx} \beta \tag{5.3.2}$$

where $\Sigma_{zy} = E[z_i y_i]$ and $\Sigma_{zx} = E[z_i x_i]$ are matrices of $K \times 1$ and $K \times L$ respectively. It is worth noticing that if $K = L$, (5.3.2) means $\beta = \Sigma_{zx}^{-1} \Sigma_{zy}$, and the model is called just-identified, where it is worth noticing that if 1 was still kept in z_i or x_i, the first row or column of Σ_{zx} would be 0, leading to the singularity of Σ_{zx}, if $K < L$, β clearly could not be solved by (5.3.2), and thus the model is non-identifiable, while if $K > L$, (5.3.2) provides more equations than the number of unknown parameters, which lead to an over-identified model.

Since it is impossible to know Σ_{zx} and Σ_{zy} in advance, the GMM substitutes them with their sample versions S_{zx} and S_{zy} respectively, where $S_{zx} = \frac{1}{N} \Sigma_{i=1}^{N} z_i x_i^T$ and $S_{zy} = \frac{1}{N} \Sigma_{i=1}^{N} z_i y_i$, then when dealing with a just-identified model, it is obvious that $\hat{\beta} = S_{zx}^T S_{zy}$, from which it is not hard to see that if $z_i = x_i$, $\hat{\beta}$ is consistent with ordinary least square estimator. The focus of

the GMM is on situation where $K > L$, where clearly there does not exist β such that (5.3.2) is satisfied completely, therefore the goal is to make $S_{zy} - S_{zx}\beta$ as close to zero as possible.

The error terms $\varepsilon_i's$ in (5.2.1) are allowed to be heteroskedastic as well as serially correlated, but in this dissertation, they are assumed to be independent. If $g_i(w_i, \beta)'s$ are also independent from each other, then it could be defined that

$$S \triangleq cov\big(g_i(w_i, \beta)\big) = E[g_i(w_i, \beta)g_i^T(w_i, \beta)] \tag{5.3.3}$$

From central limit theorem, S would be the asymptotical variance covariance matrix of $\bar{g} = \frac{1}{N}\Sigma_{i=1}^{N} g_i(w_i, \beta)$, i.e. $\bar{g} \xrightarrow{D} N(0, S)$, then given β, the sample moment estimation of S is

$$\hat{S} = \frac{1}{N}\Sigma_{i=1}^{N} g_i(w_i, \beta)g_i^T(w_i, \beta) \tag{5.3.4}$$

Let $\widehat{W}$ denote an arbitrary $K \times K$ positive definite matrix such that $\widehat{W} \xrightarrow{P} W$, where W is also positive definite, then it could be proven that

$$\hat{\beta}\big(\widehat{W}\big) = \underset{\beta}{argmin}\, N\big(S_{zy} - S_{zx}\beta\big)^T \widehat{W}^{-1}\big(S_{zy} - S_{zx}\beta\big) \tag{5.3.5}$$

has the following properties

$$\hat{\beta}\big(\widehat{W}\big) \xrightarrow{P} \beta \tag{5.3.6}$$

$$\sqrt{N}\big(\hat{\beta}\big(\widehat{W}\big) - \beta\big) \xrightarrow{d} N\Big(0, avar\big(\hat{\beta}(\widehat{W})\big)\Big) \tag{5.3.7}$$

where

$$avar\left(\hat{\beta}(\widehat{W})\right) = (\Sigma_{zx}^T W \Sigma_{zx})^{-1} \Sigma_{zx}^T W S W \Sigma_{zx} (\Sigma_{zx}^T W \Sigma_{zx})^{-1} \tag{5.3.8}$$

and a consistent estimator of $avar\left(\hat{\beta}(\widehat{W})\right)$ would be

$$\widehat{avar}\left(\hat{\beta}(\widehat{W})\right) = \left(S_{zx}^T \widehat{W} S_{zx}\right)^{-1} S_{zx}^T \widehat{W} \hat{S} \widehat{W} S_{zx} \left(S_{zx}^T \widehat{W} S_{zx}\right)^{-1} \tag{5.3.9}$$

5.4 Efficiency

Since the consistency and asymptotic normality properties could be satisfied regardless of the choice of $\widehat{W}$, a natural question to ask is: what kind of $\widehat{W}$ will generate the smallest $avar\left(\hat{\beta}(\widehat{W})\right)$, and the $\hat{\beta}\left(\widehat{W}\right)$ based on this $\widehat{W}$ will be the efficient GMM estimator.

Hansen LP (1982) showed that $\widehat{W} = \hat{S}^{-1}$, where $\hat{S} \xrightarrow{P} S$ would be the right choice. As a result, from (5.3.8) and (5.3.9),

$$avar\left(\hat{\beta}(\hat{S}^{-1})\right) = (\Sigma_{zx}^T S^{-1} \Sigma_{zx})^{-1} \tag{5.3.7}$$

$$\widehat{avar}\left(\hat{\beta}(\hat{S}^{-1})\right) = \left(S_{zx}^T \hat{S}^{-1} S_{zx}\right)^{-1} \tag{5.3.8}$$

Consequently, we are faced with a paradox that $\hat{S}^{-1}$ is needed to estimate β, while in order to obtain $\hat{S}^{-1}$, β should be known in advance, so there are the following methods dealing with this situation.

5.4.1 Two-Step Efficient GMM

Due to the fact that $\hat{\beta}(\widehat{W})$ is consistent for any arbitrary positive definite matrix $\widehat{W}$ such that $\widehat{W} \xrightarrow{P} W$, where W is also positive definite, a suitable initial choice of $\widehat{W}$ would be I_K or $(Z^T Z)^{-1}$, where Z is an $N \times K$ matrix with the i^{th} row being z_i, and $\hat{\beta}(\widehat{W})$ is the estimated β obtained from (5.3.5), then the corresponding $\hat{S}$ would be

$$\hat{S}(\widehat{W}) = \frac{1}{N} \Sigma_{i=1}^{N} z_i z_i^T \left(y_i - x_i^T \hat{\beta}(\widehat{W})\right) \tag{5.4.1}$$

Then the two-step efficient GMM estimator is

$$\hat{\beta}(\widehat{W}) = \underset{\beta}{argmin}\, N(S_{zy} - S_{zx}\beta)^T \hat{S}^{-1}(\widehat{W})(S_{zy} - S_{zx}\beta) \tag{5.4.2}$$

5.4.2 Iterated Efficient GMM

The steps indicated by (5.4.1) and (5.4.2) could be repeated until the difference between $\hat{\beta}(\widehat{W})$ from two consecutive iterations is ignorable, which in the end will generate $\hat{\beta}(\hat{S}_{iter}^{-1})$. Iterated Efficient GMM estimator and Two Step Efficient GMM estimator share the same asymptotic distribution, but the former one has the advantage of being robust the scale of data and the initial setting of $\widehat{W}$.

5.4.3 Continuous Updating Efficient GMM

Instead of estimating β iteratively like what the previous two methods do, continuous updating efficient GMM (CU) tries to estimate β and S simultaneously, which is defined as

$$\hat{\beta}\left(\hat{S}_{CU}^{-1}\right) = \underset{\beta}{argmin}\, N\left(S_{zy} - S_{zx}\beta\right)^T \hat{S}^{-1}(\beta)\left(S_{zy} - S_{zx}\beta\right) \tag{5.4.3}$$

where $\hat{S}(\beta) = \frac{1}{N}\Sigma_{i=1}^{N} z_i z_i^T (y_i - x_i^T\beta)^2$. CU has the same merit as iterated efficient GMM but is burdensome to compute, while the finite sample performance of it is superior to the other two.

5.5 Model Checking

5.5.1 J-Statistic

The J-Statistic is used to test whether the orthogonal conditions indicated by (5.3.1) is valid, and it is defined as

$$J = J\left(\hat{\beta}\left(\hat{S}^{-1}\right), \hat{S}^{-1}\right) = N\left(S_{zy} - S_{zx}\hat{\beta}\left(\hat{S}^{-1}\right)\right)^T \hat{S}^{-1}\left(S_{zy} - S_{zx}\hat{\beta}\left(\hat{S}^{-1}\right)\right) \tag{5.5.1}$$

For just-identified model, i.e. $K = L$, J is always zero, while for over-identified model with $K > L$, which is often the case, then under H_0: (5.3.1) is satisfied, one should expect $J \xrightarrow{d} \chi^2_{K-L}$. Hence J-statistic is a general test of modeling setting, and a large one indicates model mis-specification, however, it cannot provide information about how the model is mis-specified.

5.5.2 Normalized Moments

If the model is rejected by the J-statistic, then it would be of interest to locate the source of this rejection, which could be indicated by the normalized moments $\sqrt{N}\left(S_{zy} - S_{zx}\hat{\beta}\left(\hat{S}^{-1}\right)\right)$, because under the null hypothesis, that is the model is correct and the orthogonal conditions are satisfied, we have

$$\sqrt{N}\left(S_{zy} - S_{zx}\hat{\beta}\left(\hat{S}^{-1}\right)\right) \xrightarrow{D} N(0, S - \Sigma_{zx}[\Sigma_{zx}^{T}S^{-1}\Sigma_{zx}]^{-1}\Sigma_{zx}^{T}) \tag{4.5.3}$$

As a result, the individual moment t-ratio

$$t_i = \frac{\left(\left(S_{zy} - S_{zx}\hat{\beta}(\hat{S}^{-1})\right)\right)_i}{\sqrt{\left(\hat{S} - \mathrm{S}_{zx}[\mathrm{S}_{zx}^{T}\hat{S}^{-1}\mathrm{S}_{zx}]^{-1}\mathrm{S}_{zx}^{T}\right)_{ii}/T}} \tag{5.5.4}$$

is asymptotically standard normal, thus a large t_i indicates the mis-specification of the i^{th} orthogonal condition.

Chapter 6. Literature reviews of platform comparison methods

6.1 Introduction

It is quite common in any discipline that certain concepts could be measured by multiple techniques, e.g., gene expression level could be measured by microarray, next generation sequencing or qPCR etc. Due to the fact that these latent concepts could not be observed directly, SEM seems to be a quite suitable model to analyze multiple platforms for a certain concept.

Surprisingly there are not too many literatures on using SEM to perform platform comparison, and related works include Sun et al (2014) who applied SEM to calibrate qPCR, microarray and RNA-sequencing (RNA-seq) and further estimated the true expression level of each gene [19], and the same group also published a paper where SEM was used to compare different normalization methods of RNA-seq [20]. Besides SEM, conventional methods like Pearson Correlation among different platforms, reproducibility within each platform, are often used a criteria of platform quality, e.g. Spurgeon et al (2008) [21], Chen et al (2007) [22], Arikawa et al (2008) [23] all used similar methods to compare multiple gene expression measurement methods. However, these conventional criteria have been suffering from critics since they are not sophisticated enough to capture the information of agreement among platforms, and regression based models are in demand to handle the task. Allen et al (1997) applied both Pearson Correlation and ordinary least squares to compare among different techniques of measuring density of ambient particulate matters [24]. While as discussed in

Chapter 4, OLS is not suitable in this situation since each platforms is subject to measurement error, thus more advanced models are in need.

This chapter mainly focuses on the work done by Xiao Wu et al (2013) [25], where she adopted the latent SEM to compare multiple platforms measuring the abundance of bacteria, including Sanger sequencing, next generation pyrosequencing with two windows (454_V1V3 and 454_V3V5), and quantitative PCR (qPCR), and further identified the most reliable platform. The contents in the next chapter were actually motivated by her work since she modeled each taxon of bacterium separately, and thus the results differ across different bacterium, which is why the random effects are adopted in order to perform an overall comparison while allowing individual (bacterium) heterogeneity. At the end of this chapter, another important work of applying SEM on platform comparison will also be reviewed.

6.2 Data Structure

ABI 3730 Sanger sequencing [26] and 454 FLX Titanium pyrosequencing [27] including two hypervariable regions V1V3 and V3V5, which belongs to the next generation sequencing (NGS) technology, were used to generate data from 300 healthy human subjects by amplifying 16S rRNA genes. In addition, quantitative polymerase chain reaction (qPCR) [28], which employs primers to detect and quantify bacteria, are also available for a single bacterial taxon, Faecalibacterium spp.

Besides Faecalibacterium, measurements of several other bacteria including Proteobacteria, Firmicutes/Clostridia/Clostridiales/LachnolV, Actinobacteria, Bacteroidetes, Firmicutes/Bacilli are also available in Sanger, 454_V1V3 and 454_V3V5.

6.3 Model Setting

For each bacterium, the true frequency of subject i is considered as a latent variable ξ_i, while the corresponding measurements from p platforms are denoted as $X = \left(X_{i1}, \cdots, X_{ip}\right)^T$, which are observable, then based on the model setting in secton 2.2, it follows naturally that $X_{ij} = \lambda_j \xi_i + e_{ij}$, where $var(\xi_i) = 1$.

Given normality assumption of ξ and $\varepsilon = \left(\varepsilon_1, \cdots, \varepsilon_p\right)$, i.e. $\xi \sim N(0,1)$ and $\varepsilon \sim MVN\left(0, \sigma^2 I_p\right)$, it could be obtained that $X \sim MVN\left(0, \Lambda\Lambda^T + \sigma^2 I_p\right)$, where $\Lambda = \left(\lambda_1, \cdots, \lambda_p\right)^T$. then the log likelihood becomes

$$l \propto -\frac{N}{2} log\left|\Lambda\Lambda^T + \sigma^2 I_p\right| - \frac{1}{2}\Sigma_{j=1}^{N} X_j^T \left(\Lambda\Lambda^T + \sigma^2 I_p\right)^{-1} X_j \qquad (6.3.1)$$

from which the maximum likelihood estimates could be obtained.

In terms of platform quality, it is natural to use reliability as an index, which is defined as

$$R_{X_i}^2 = \frac{var(\lambda_i \xi)}{var(X_i)} = 1 - \frac{var(\varepsilon_i)}{var(X_i)} \qquad (6.3.2)$$

i.e. the percentage of the variance of X_i that is explained by the model.

The process described above could be used to compare Sanger, 454_V1V3, 454_V3V5, qPCR for *Faecalibacterium*, and compare Sanger, 454_V1V3, 454_V3V5 for all the other bacteria, the diagrams of which are indicated by Figure 6.1.

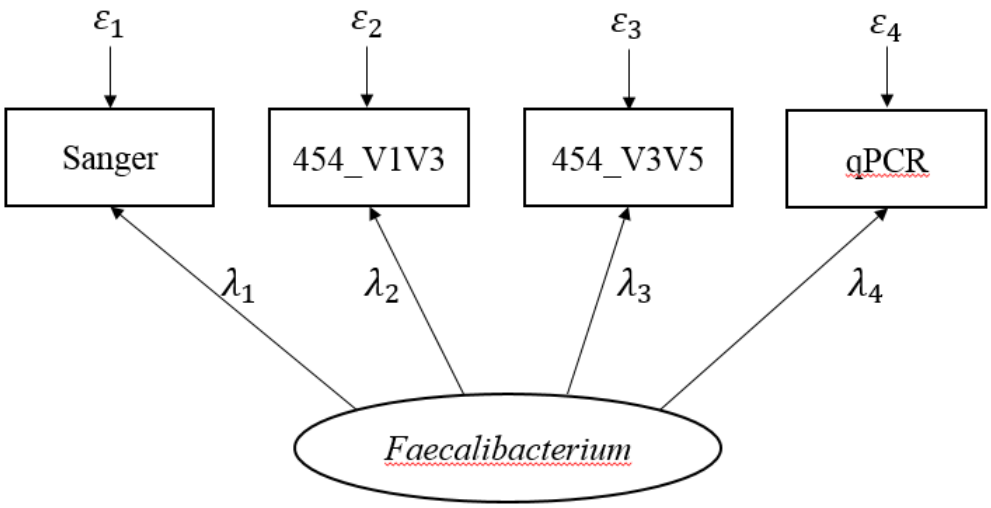

Figure 6.1. SEM comparing measurements of abundance of *Faecalibacterium* from Sanger, 454_V1V3, 454V3V5 and qPCR.

6.4 Results

The result of comparison among Sanger, 454_V1V3, 454_V3V5 and qPCR for *Faecalibacterium* is shown in Figure 6.2, where 454_V3V5 has the highest loading, 0.955, and the reliabilities of these four platforms, computed by (6.3.2), are 0.819, 0.857, 0.912 and 0.441 respectively.

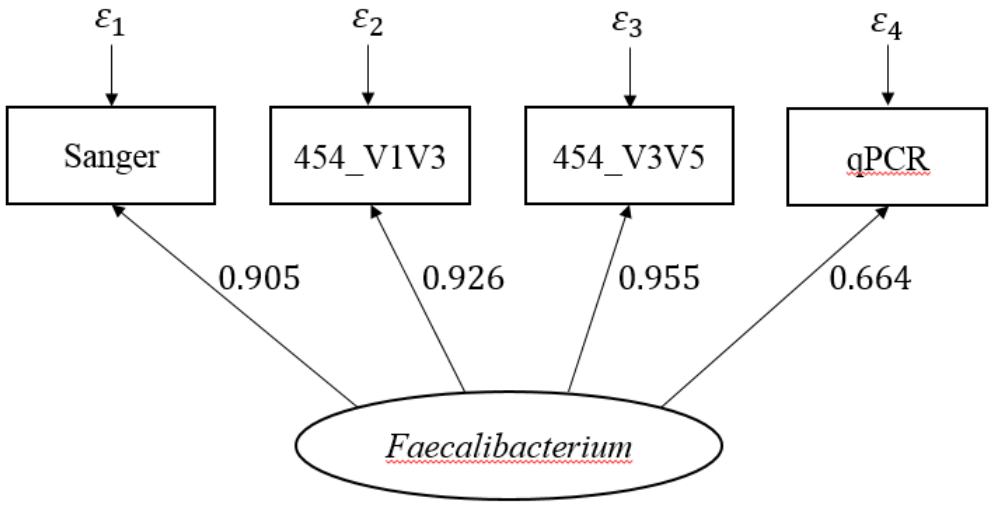

Figure 6.2. Estimation results of SEM comparing measurements of abundance of *Faecalibacterium* from Sanger, 454_V1V3, 454V3V5 and qPCR.

As mentioned in Section 6.2, measurements from Sanger, 454_V1V3 and 454_V3V5 are available for other bacterium, thus similar analysis could be done on each one of them, whose reliabilities are shown in Table 6.1.

Table 6.1. Reliabilities of Sanger, 454_V1V3 and 454_V3V5 when comparing measurements of abundance of Proteobacteria, Firmicutes/Clostridia/Clostridiales/LachnoIV, Actinobacteria, Bacteroidetes and Firmicutes/Bacilli.

Reliability	Sanger	454_V1V3	454_V3V5
Proteobacteria	0.657	0.641	0.974
Firmicutes/Clostridia/Clostridiales/LachnoIV	0.685	0.923	0.793
Actinobacteria	0.582	0.854	0.882
Bacteroidetes	0.684	0.828	0.980
Firmicutes/Bacilli	0.698	0.953	0.959

6.5 Discussion

Both Figure 6.2 and Table 6.1 show that for most bacterium, 454_V3V5 is superior than others, but for Firmicutes/Clostridia/Clostridiales/LachnoIV, 454_V1V3 performs the best, then the same issue as in Section 4.6.3 occurs, meaning it is not reasonable to assume that platforms perform homogeneously across different bacteria. Therefore, it is of our interest to know whether the platforms are consistent or not, or which one performs the best in general, while at the same time, the behavior of platforms should be allowed to vary across bacteria. Therefore, a model that could handle this issue will be introduced in Chapter 7.

6.6 Another related work

Bilonick et al (2015) proposed the framework of comparing multiple samplers of measuring density of PM2.5 using linked structural equation modeling [29]. In this work, three federal references methods (FRM1, FRM2, FRM3), three speciation samplers (SASS, SFS, IMP), and a tapered element oscillating microbalance (TEOM) were compared in terms of measuring PM2.5, and furthermore, calibration between each pair under different temperatures were also established. To stabilize the variance, square root data was analyzed instead of raw data.

Figure 6.3 illustrates eight sub-SEM models comparing seven platforms under eight temperatures that are -5.8°C, 0.7°C, 5.3°C, 10.0°C, 14.4°C, 18.1°C, 21.1°C and 24.4°C, and

preliminary knowledge about samplers indicates that only TEOM is affected by temperatures, which is the reason that only its loadings vary across eight models.

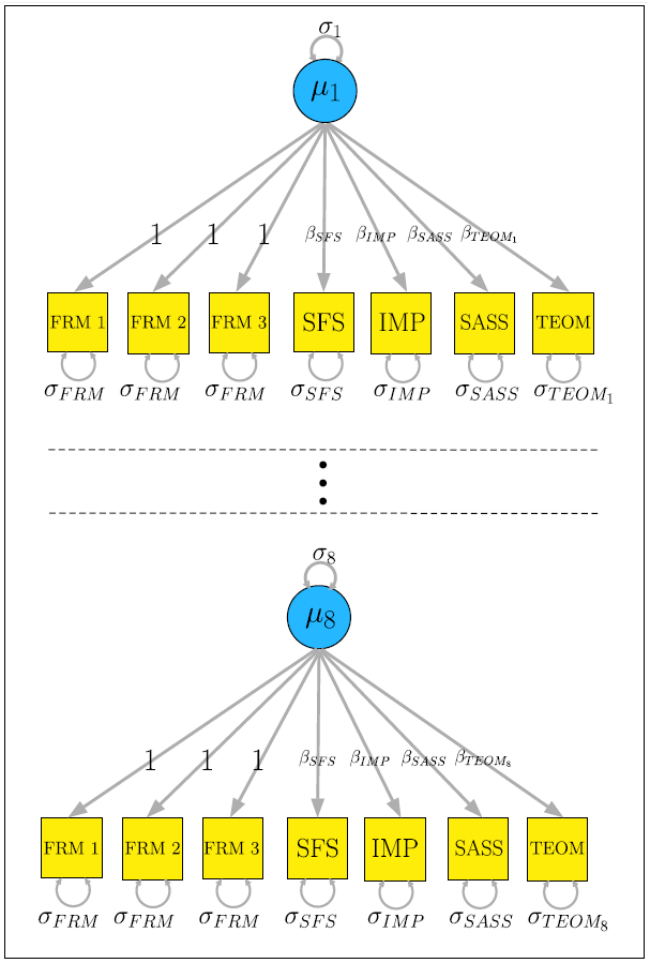

Figure 6.3. From Bilonick et al (2015), path diagram for structural equation model for measurement error relating all seven samplers and stratified by temperature. This model is composed of eight submodels with some parameters constrained to be equal across the temperature strata.

FRM1, FRM2, FRM3 are three identical samplers whose measurements were took on different frequencies of days, thus they could be considered as technical replicates, and their loadings are equal to each other as shown in Figure 6.3.

After fitting eight SEMs together, eight estimates of $\hat{\alpha}_{TEOM_t}$, $\hat{\beta}_{TEOM_t}$ and $\frac{\hat{\sigma}_{TEOM_t}}{\hat{\beta}_{TEOM_t}}$ for $t = 1, \cdots, 8$ could be obtained, and their scatterplot versus temperatures are shown in Figure 6.4A, B and C. Due to the sigmoid shape of $\hat{\alpha}_{TEOM_t}$, $\hat{\beta}_{TEOM_t}$ and the linear shape of $\frac{\hat{\sigma}_{TEOM_t}}{\hat{\beta}_{TEOM_t}}$, it was assumed that $\hat{\alpha}_{TEOM_t} = A_\alpha + \frac{B_\alpha - A_\alpha}{1+e^{\frac{t_\alpha - t}{S_\alpha}}}$, $\hat{\beta}_{TEOM_t} = A_\beta + \frac{B_\beta - A_\beta}{1+e^{\frac{t_\beta - t}{S_\beta}}}$ and $\frac{\hat{\sigma}_{TEOM_t}}{\hat{\beta}_{TEOM_t}} = a + bt$ for any temperature t, and the model is re-fitted based on these shape assumptions.

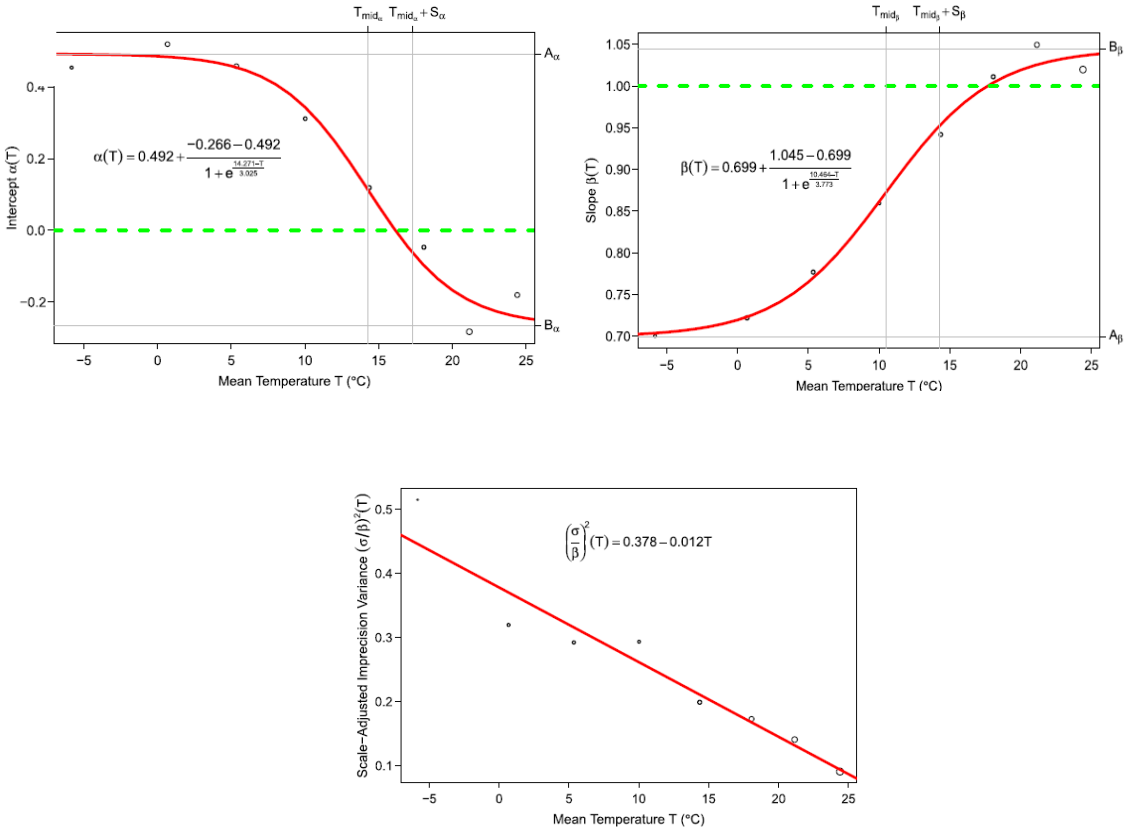

Figure 6.4. From Bilonick et al (2015), A – Fitted sigmoid function of $\hat{\alpha}_{TEOM_t}$, B – Fitted sigmoid function of $\hat{\beta}_{TEOM_t}$, C – Fitted linear function of $\frac{\hat{\sigma}_{TEOM_t}}{\hat{\beta}_{TEOM_t}}$.

After finalizing the estimates, one could easily obtain the calibration relation between any pair of samplers. For example, with FRM and TEOM at temperature t, it could be obtained that $\sqrt{FRM} = \mu + \varepsilon_{FRM}$ and $\sqrt{TEOM} = \hat{\alpha}_{TEOM_t} + \hat{\beta}_{TEOM_t}\mu + \varepsilon_{TEOM}$, where μ is the true PM2.5 density, ε_{FRM} and ε_{TEOM} are the corresponding residuals of two samplers, then it follows naturally that $TEOM = \left(\hat{\alpha}_{TEOM_t} + \hat{\beta}_{TEOM_t}\sqrt{FRM}\right)^2$, which was shown to be more plausible than fitting OLS between $\sqrt{TEOM}$ and $\sqrt{FRM}$ in their paper.

An important contribution of their work to this dissertation is, given preliminary knowledge of all the platforms, sub-SEM models could be linked by constraining some parameters to be identical across strata while allowing others to vary, which is why in Chapter 7, all of the loadings of each platform across different strata will be assumed to consist of a mean loading and a random effect to cope with situations where there are no available preliminary knowledge on platforms.

Chapter 7. SEM and EIV with Random Effects

7.1 Introduction

Multiple measurement platforms of Microbiome abundance are increasingly available nowadays, including microarray, next-generation sequencing, quantitative PCR etc., thus the evaluation of the consistency of which, has become an increasingly urgent topic. Existing methods including using Pearson correlation or the EIV to gauge the linear dependency between two platforms [13], applying structural equation modeling (SEM) to estimate the relations among three or more platforms [25] etc., are mainly designed to determine the agreement of platforms on each individual bacterium without taking into account the heterogeneity of individual bacterium to yield an overall platform agreement measure across the entire Microbiome. Reasons that such heterogeneity should be considered have been covered at the end of Chapter 4 and Chapter 6.

In this work, we develop a novel method for overall platform agreement analysis via SEM or EIV via the random effect model. Our method is illustrated through a 16S ribosomal RNA sequencing study measuring bacteria abundance via three measurements windows: V1V2, V1V3 and V3V4. We found good agreement between V1V2 and V3V4, and between V1V3 and V3V4 is found, however, more discrepancy was found between V1V2 and V1V3 with p value of $2.4e-7$, which strongly rejected the null hypothesis that they were consistent. Moreover, the

prediction of random loadings, a by-product of the model above, is able to elucidate the performance of platforms on each individual bacterium.

The paradigm mentioned above could be easily adjusted to situations where only two platforms are available via the Errors in variables (EIV) model, which is another contribution of this work. To further confirm the conclusions above, pairwise comparison is performed and we are glad to report the random effect SEM and the random effect EIV model yielded consistent results.

7.2 Background

16S ribosomal RNA (rRNA) sequencing has been a well-established method of profiling amplicons to identify and enumerate bacteria present in a given sample due to merits including its presence in almost all bacteria, stable function over time and large bp size for informatic purposes [30]. There are nine hypervariable amplicon regions targeted in the 16S gene, i.e. V1 to V9 [31], of which three were selected in this study for the check of consistency, which are V1V2, V1V3 and V3V4, hereafter refered to as three platforms, and it is of our interest to compare the consistency among them.

Due to the multiple options of targeting regions, it is of major interest to study the consistency among measurements resulting from all of them. Instead of treating this consistency as a fixed property across all bacteria, which is not uncommon when people did platform

comparison, e.g. in this study, we considered that property as random across different bacteria. Consequently, the mean of that random consistency, i.e. the fixed effect part, served as the criteria of consistency between regions, or platforms in a broader sense, while the consistency for each bacteria.

7.3 Data Structure

240 bacteria were measured on the same 6 rats in each platform, with each rat repeated 10 times. In order for the raw counts to be comparable across platforms, measurements were transformed into percentage by dividing each count by the total count of all bacteria of that rat. In additional, bacteria with percentages of all of the replicates from all of the rats equal to 0 in any one of the three platforms were filtered out, which led to 55 bacteria left.

To make the measurements more normally distributed and to stabilize the variance, arcsine square root transformation [32] was applied on the percentages, where each p would be transformed into $arcsin\left(\sqrt{p}\right)$. Moreover, if a certain percentage is 0, it would be transformed into $arcsin\left(\frac{1}{4n}\right)$, where n is the total counts of all of the bacteria for that particular replicate of rat. The data structure of measurements from V1V2 (X) is indicated by Table 7.1, and V1V3 (Y) and V3V4 (Z) follow the same pattern.

Table 7.1. Data structure of measurements form V1V2.

V1V2	Rat 1			…	Rat 6		
	R1	…	R10		R1	…	R10
Bacteria/Acidobacteria/Acidobacteria /.../Edaphobacter	X_{11}^{1}	…	X_{11}^{10}	…	X_{16}^{1}	…	X_{16}^{10}
Bacteria/Acidobacteria/Acidobacteria /.../AKIW659	X_{21}^{1}	…	X_{21}^{10}	…	X_{26}^{1}	…	X_{26}^{10}
⋮		⋮		⋱		⋮	
Bacteria/Verrucomicrobia/Verrucomicrobiae /.../Akkermansia	X_{I1}^{1}	…	X_{I1}^{10}	…	X_{I6}^{1}	…	X_{I6}^{10}

7.4 Model Setting

In each platform, it is assumed that for each bacterium, even if it does not exist, there will be an unknown, but fixed non-zero measurement, which is called constant systematic error [16] for this platform. These errors are defined as α_0, β_0 and γ_0 respectively for V1V2 (X), V1V3 (Y) and V3V4 (Z).

The true abundance of i^{th} bacterium from the j^{th} subject is considered as a unobservable latent variable ξ_{ij}, which satisfies $\xi_{ij} \sim N(\xi_i, \sigma_{\xi_i}^2)$, then the corresponding measurement, e.g. from V1V2 (X) could be affected by a factor of α_1, which is called proportional systematic error [16]. In order to incorporate the heterogeneity of each bacterium i, a random effect a_{i1}, is added to α_1, which gives $A_{i1} = \alpha_1 + a_{i1}$. In parallel, there are $B_{i1} = \beta_1 + b_{i1}$ and $C_{i1} = \gamma_1 + c_{i1}$ for V1V3 (Y) and V3V4 (Z) respectively. Therefore, the measurements from three platforms are modeled as

$$\begin{cases} X_{ij}^k = \alpha_0 + A_{i1}\xi_{ij} + \delta_{ij}^k \\ Y_{ij}^k = \beta_0 + B_{i1}\xi_{ij} + \varepsilon_{ij}^k \\ Z_{ij}^k = \gamma_0 + C_{i1}\xi_{ij} + \tau_{ij}^k \end{cases} \tag{7.4.1}$$

where $i = 1, \cdots, I$, $j = 1, \cdots, J$, $k = 1, \cdots, K$, and $I = 55$ is the number of bacteria, $J = 6$ is the number of rats, $K = 10$ is the number of replicates per rat. Besides, X_{ij}^k is the measurement of i^{th} bacterium from the k^{th} replicate of the j^{th} subject in terms of V1V2, then Y_{ij}^k and Z_{ij}^k are the counterparts of V1V3 and V3V4 respectively. Figure 7.1 is the diagram of model defined by (7.4.1).

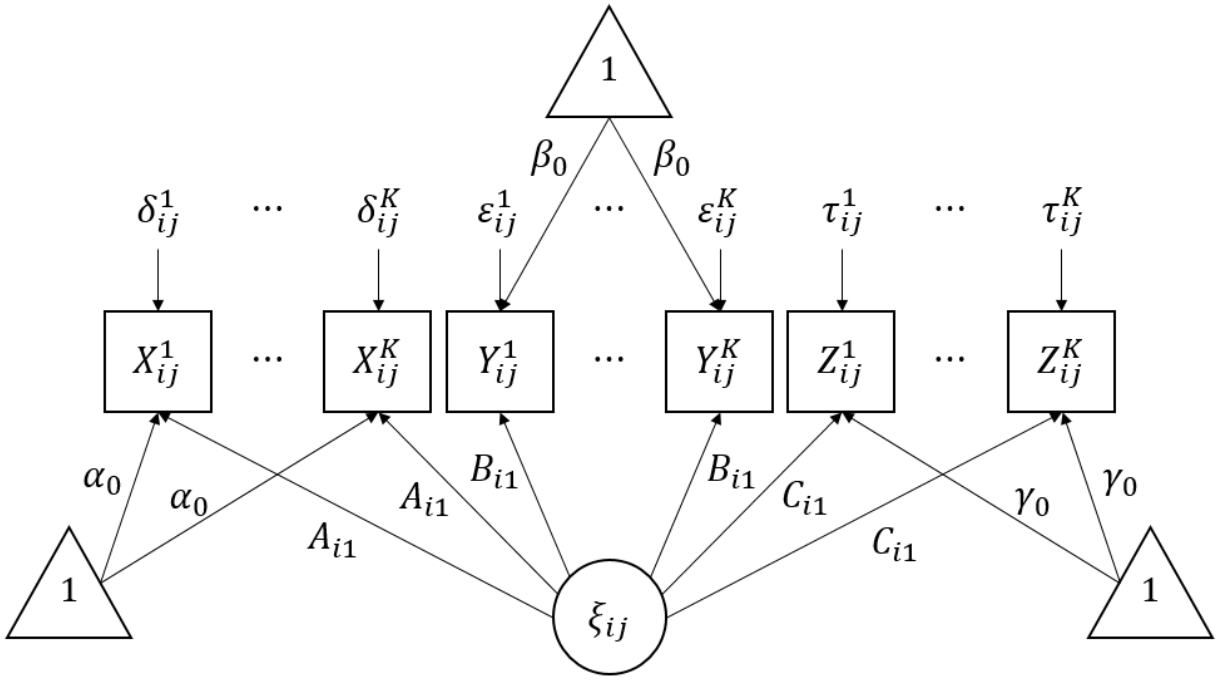

Figure 7.1. Diagram of model defined by Equation (7.4.1), which is SEM with random effects.

Normality assumptions of residuals in each platform are made for the purpose of model fitting, which include $\delta_{ij}^k \sim N(0, \sigma_{\delta_i}^2)$, $\varepsilon_{ij}^k \sim N(0, \sigma_{\varepsilon_i}^2)$ and $\tau_{ij}^k \sim N(0, \sigma_{\tau_i}^2)$. For the sake of model

identifiability, constraints need to be put on α_1, β_1 or γ_1. Without loss of generality, α_1 is constrained to be 1, meaning that V1V2 serves as a reference platform against which V1V3 and V3V4 would be compared. As a result, $A_{i1}\sim N(1,\sigma_{A_1}^2)$, $B_{i1}\sim N(\beta_1,\sigma_{B_1}^2)$ and $C_{i1}\sim N(\gamma_1,\sigma_{C_1}^2)$.

Given the model settings above, if denoting $X_{ij}=\left(X_{ij}^1,\cdots,X_{ij}^K\right)^T$, $X_i=\left(X_{i1}^T,\cdots,X_{iJ}^T\right)^T$, and similarly for Y_i and Z_i, then with the definitions below, namely I_p is the p dimensional identity matrix, E_p is the p dimensional square matrix with all elements equal to 1, and $diag_p(M)$ is the block diagonal matrix with M at the diagonal positions repeatedly for p times, it follows naturally that

$$D_i \triangleq (X_i^T, Y_i^T, Z_i^T)^T \sim N(\mu_i, V_i) \tag{7.4.2}$$

where

$$\mu_i = \left[(\alpha_0+\xi_i)1_{JK}^T, (\beta_0+\beta_1\xi_i)1_{JK}^T, (\gamma_0+\gamma_1\xi_i)1_{JK}^T\right]^T \tag{7.4.3}$$

$$V_i = \begin{bmatrix} V_{i1} & V_{i12} & V_{i13} \\ V_{i12}^T & V_{i2} & V_{i23} \\ V_{i13}^T & V_{i23}^T & V_{i3} \end{bmatrix} \tag{7.4.4}$$

$$V_{i1} \triangleq VAR(X_i) = \sigma_{\delta_i}^2 I_{JK} + \xi_i^2\sigma_{A_1}^2 E_{JK} + diag_J\left(\left(1+\sigma_{A_1}^2\right)\sigma_{\xi_i}^2 I_K\right) \tag{7.4.5}$$

$$V_{i2} \triangleq VAR(Y_i) = \sigma_{\varepsilon_i}^2 I_{JK} + \xi_i^2\sigma_{B_1}^2 E_{JK} + diag_J\left(\left(1+\sigma_{B_1}^2\right)\sigma_{\xi_i}^2 I_K\right) \tag{7.4.6}$$

$$V_{i3} \triangleq VAR(Z_i) = \sigma_{\tau_i}^2 I_{JK} + \xi_i^2\sigma_{C_1}^2 E_{JK} + diag_J\left(\left(1+\sigma_{C_1}^2\right)\sigma_{\xi_i}^2 I_K\right) \tag{7.4.7}$$

$$V_{i12} \triangleq COV(X_i, Y_i) = diag_J\left(\beta_1\sigma_{\xi_i}^2 I_K\right) \tag{7.4.8}$$

$$V_{i13} \triangleq COV(X_i, Z_i) = diag_J(\gamma_1 \sigma_{\xi_i}^2 I_K) \tag{7.4.9}$$

$$V_{i23} \triangleq COV(Y_i, Z_i) = diag_J(\beta_1 \gamma_1 \sigma_{\xi_i}^2 I_K) \tag{7.4.10}$$

thus it could be obtained that the log likelihood function of all observations satisfies

$$l \propto -\frac{1}{2}\Sigma_{i=1}^{I}[log|V_i| + (D_i - \mu_i)^T V_i^{-1}(D_i - \mu_i)] \tag{7.4.11}$$

From the model settings above, the process of data preparation and adopting random effects could be depicted by Figure 7.2 below.

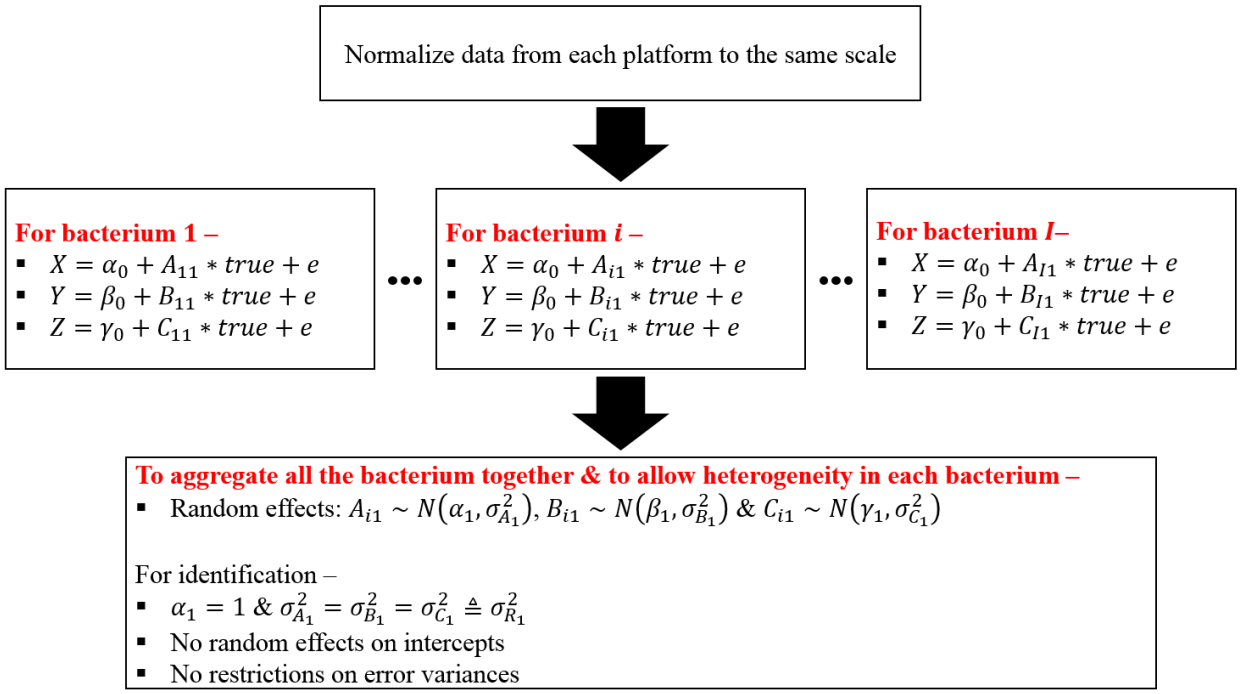

Figure 7.2. Flowchart of SEM with random effects based on the model setting in Section 7.4.

7.5 Estimation

EM algorithm [5] is adopted here since the MLE of the original likelihood is cumbersome to solve, where

$$\Theta \triangleq \left(\alpha_0, \beta_0, \gamma_0, \beta_1, \gamma_1, \sigma_{A_1}^2, \sigma_{B_1}^2, \sigma_{C_1}^2, \xi_i, \sigma_{\xi_1}^2, \sigma_{\delta_i}^2, \sigma_{\varepsilon_i}^2, \sigma_{\tau_i}^2\right) \tag{7.5.1}$$

is the vector containing all of the parameters to be estimated, and

$$\Lambda \triangleq \left(A_{i1}, B_{i1}, C_{i1}, \xi_{ij}, \delta_{ij}^k, \varepsilon_{ij}^k, \tau_{ij}^k\right) \tag{7.5.2}$$

is the vector containing all of the unobserved missing variables, then the log likelihood of the completed data would be

$$\begin{aligned} l_c \propto & -\frac{3I}{2} log\sigma_{R_1}^2 - \frac{1}{2}\Sigma_{i=1}^{I}\left[\frac{(A_{i1}-1)^2}{\sigma_{R_1}^2} + \frac{(B_{i1}-1)^2}{\sigma_{R_1}^2} + \frac{(C_{i1}-1)^2}{\sigma_{R_1}^2}\right] \\ & -\frac{1}{2}\Sigma_{i=1}^{I}\left[Jlog\sigma_{\xi_i}^2 + \Sigma_{j=1}^{J}\frac{\left(\xi_{ij}-\xi_i\right)^2}{\sigma_{\xi_i}^2}\right] - \frac{JK}{2}\Sigma_{i=1}^{I}\left[log\sigma_{\delta_i}^2 + log\sigma_{\varepsilon_i}^2 + log\sigma_{\tau_i}^2\right] \\ & -\frac{1}{2}\Sigma_{i=1}^{I}\Sigma_{j=1}^{J}\Sigma_{k=1}^{K}\left[\frac{\left(\delta_{ij}^k\right)^2}{\sigma_{\delta_i}^2} + \frac{\left(\varepsilon_{ij}^k\right)^2}{\sigma_{\varepsilon_i}^2} + \frac{\left(\tau_{ij}^k\right)^2}{\sigma_{\tau_i}^2}\right] \end{aligned} \tag{7.5.3}$$

The EM algorithm consists of the Expectation Step (E step) and the Maximization Step (M step). At the E step, conditional expectation of l_c given observations $D_i, i = 1, \cdots, I$ and the current parameter estimation $\hat{\Theta}^{(t)}$, i.e. $E\left[l_c | D_i, \hat{\Theta}^{(t)}\right]$, is obtained.

With the following definitions and derivations,

$$\delta_{ij} \triangleq (\delta_{ij}^1, \cdots, \delta_{ij}^K) \text{ and } \delta_i \triangleq \left(\delta_{i1}^T, \cdots, \delta_{iJ}^T\right)^T \tag{7.5.4}$$

$$\varepsilon_{ij} \triangleq (\varepsilon_{ij}^1, \cdots, \varepsilon_{ij}^K) \text{ and } \varepsilon_i \triangleq \left(\varepsilon_{i1}^T, \cdots, \varepsilon_{iJ}^T\right)^T \tag{7.5.5}$$

$$\tau_{ij} \triangleq (\tau_{ij}^1, \cdots, \tau_{ij}^K) \text{ and } \tau_i \triangleq \left(\tau_{i1}^T, \cdots, \tau_{iJ}^T\right)^T \tag{7.5.6}$$

$$V_{\delta_i} \triangleq VAR(\delta_i) = \sigma_{\delta_i}^2 I_{JK}, V_{\varepsilon_i} \triangleq VAR(\varepsilon_i) = \sigma_{v_i}^2 I_{JK}, \text{ and } V_{\varepsilon_i} \triangleq VAR(\varepsilon_i) = \sigma_{v_i}^2 I_{JK} \tag{7.5.7}$$

$$\vec{\xi}_i \triangleq \left(\xi_{i1}, \cdots, \xi_{iJ}\right)^T \text{ and } V_{\vec{\xi}_i} \triangleq VAR\left(\vec{\xi}_i\right) = \sigma_{\xi_i}^2 I_J \tag{7.5.8}$$

$$V_{\vec{\xi}_i, D_i} \triangleq COV\left(\vec{\xi}_i, D_i\right) = \left[\sigma_{\xi_i}^2 E_{JK}, \beta_1 \sigma_{\xi_i}^2 E_{JK}, \gamma_1 \sigma_{\xi_i}^2 E_{JK}\right] \tag{7.5.9}$$

$$V_{\delta_i, D_i} \triangleq COV(\delta_i, D_i) = \left[V_{\delta_i}, 0_{JK \times JK}, 0_{JK \times JK}\right] \tag{7.5.10}$$

$$V_{\varepsilon_i, D_i} \triangleq COV(\varepsilon_i, D_i) = \left[0_{JK \times JK}, V_{\varepsilon_i}, 0_{JK \times JK}\right] \tag{7.5.11}$$

$$V_{\tau_i, D_i} \triangleq COV(\tau_i, D_i) = \left[0_{JK \times JK}, 0_{JK \times JK}, V_{\tau_i}\right] \tag{7.5.12}$$

$$V_{A_{i1}, D_i} \triangleq COV(A_{i1}, D_i) = \left[\xi_i \sigma_{R_1}^2, 0_{1 \times JK}, 0_{1 \times JK}\right] \tag{7.5.13}$$

$$V_{B_{i1}, D_i} \triangleq COV(B_{i1}, D_i) = \left[0_{1 \times JK}, \xi_i \sigma_{R_1}^2, 0_{1 \times JK}\right] \tag{7.5.14}$$

$$V_{C_{i1}, D_i} \triangleq COV(C_{i1}, D_i) = \left[0_{1 \times JK}, 0_{1 \times JK}, \xi_i \sigma_{R_1}^2\right] \tag{7.5.15}$$

it could be obtained that

$$\tilde{\delta}_i^{(t)} \triangleq E\left[\delta_i | D_i, \hat{\theta}^{(t)}\right] = \hat{V}_{\delta_i, D_i}^{(t)} \cdot \left[\hat{V}_i^{(t)}\right]^{-1} \cdot \left[D_i - \hat{\mu}_i^{(t)}\right] \tag{7.5.16}$$

$$\tilde{\sigma}_{\delta_i}^{2(t)} \triangleq VAR(\delta_i | D_i, \hat{\theta}^{(t)}) = \hat{V}_{\delta_i}^{(t)} - \hat{V}_{\delta_i, D_i}^{(t)} \cdot \left[\hat{V}_i^{(t)}\right]^{-1} \cdot \left[\hat{V}_{\delta_i, D_i}^{(t)}\right]^T \tag{7.5.17}$$

$$\tilde{\varepsilon}_i^{(t)} \triangleq E\left[\varepsilon_i | D_i, \hat{\Theta}^{(t)}\right] = \hat{V}_{\varepsilon_i, D_i}^{(t)} \cdot \left[\hat{V}_i^{(t)}\right]^{-1} \cdot \left[D_i - \hat{\mu}_i^{(t)}\right] \tag{7.5.18}$$

$$\tilde{\sigma}_{\varepsilon_i}^{2(t)} \triangleq VAR(\varepsilon_i | D_i, \hat{\Theta}^{(t)}) = \hat{V}_{\varepsilon_i}^{(t)} - \hat{V}_{\varepsilon_i, D_i}^{(t)} \cdot \left[\hat{V}_i^{(t)}\right]^{-1} \cdot \left[\hat{V}_{\varepsilon_i, D_i}^{(t)}\right]^T \tag{7.5.19}$$

$$\tilde{\tau}_i^{(t)} \triangleq E\left[\tau_i | D_i, \hat{\Theta}^{(t)}\right] = \hat{V}_{\tau_i, D_i}^{(t)} \cdot \left[\hat{V}_i^{(t)}\right]^{-1} \cdot \left[D_i - \hat{\mu}_i^{(t)}\right] \tag{7.5.20}$$

$$\tilde{\sigma}_{\tau_i}^{2(t)} \triangleq VAR(\tau_i | D_i, \hat{\Theta}^{(t)}) = \hat{V}_{\tau_i}^{(t)} - \hat{V}_{\tau_i, D_i}^{(t)} \cdot \left[\hat{V}_i^{(t)}\right]^{-1} \cdot \left[\hat{V}_{\tau_i, D_i}^{(t)}\right]^T \tag{7.5.21}$$

$$\tilde{\xi}_i^{(t)} \triangleq E\left[\vec{\xi}_i | D_i, \hat{\Theta}^{(t)}\right] = \hat{V}_{\vec{\xi}_i, D_i}^{(t)} \cdot \left[\hat{V}_i^{(t)}\right]^{-1} \cdot \left[D_i - \hat{\mu}_i^{(t)}\right] \tag{7.5.22}$$

$$\tilde{\sigma}_{\xi_i}^{2(t)} \triangleq VAR\left(\vec{\xi}_i | D_i, \hat{\Theta}^{(t)}\right) = \hat{V}_{\vec{\xi}_i}^{(t)} - \hat{V}_{\vec{\xi}_i, D_i}^{(t)} \cdot \left[\hat{V}_i^{(t)}\right]^{-1} \cdot \left[\hat{V}_{\vec{\xi}_i, D_i}^{(t)}\right]^T \tag{7.5.23}$$

$$\tilde{A}_{i1}^{(t)} \triangleq E\left[A_{i1} | D_i, \hat{\Theta}^{(t)}\right] = 1 + \hat{V}_{A_{i1}, D_i}^{(t)} \cdot \left[\hat{V}_i^{(t)}\right]^{-1} \cdot \left[D_i - \hat{\mu}_i^{(t)}\right] \tag{7.5.24}$$

$$\tilde{\sigma}_{A_1}^{2(t)} \triangleq VAR\left(A_{i1} \middle| D_i, \hat{\Theta}^{(t)}\right) = \hat{\sigma}_{A_1}^{2(t)} + \hat{V}_{A_{i1}, D_i}^{(t)} \cdot \left[\hat{V}_i^{(t)}\right]^{-1} \cdot \left[\hat{V}_{A_{i1}, D_i}^{(t)}\right]^T \tag{7.5.25}$$

$$\tilde{B}_{i1}^{(t)} \triangleq E\left[B_{i1} | D_i, \hat{\Theta}^{(t)}\right] = \hat{\beta}_1^{(t)} + \hat{V}_{B_{i1}, D_i}^{(t)} \cdot \left[\hat{V}_i^{(t)}\right]^{-1} \cdot \left[D_i - \hat{\mu}_i^{(t)}\right] \tag{7.5.26}$$

$$\tilde{\sigma}_{B_1}^{2(t)} \triangleq VAR\left(B_{i1} \middle| D_i, \hat{\Theta}^{(t)}\right) = \hat{\sigma}_{R_1}^{2(t)} + \hat{V}_{B_{i1}, D_i}^{(t)} \cdot \left[\hat{V}_i^{(t)}\right]^{-1} \cdot \left[\hat{V}_{B_{i1}, D_i}^{(t)}\right]^T \tag{7.5.27}$$

$$\tilde{C}_{i1}^{(t)} \triangleq E\left[C_{i1} | D_i, \hat{\Theta}^{(t)}\right] = \hat{\gamma}_1^{(t)} + \hat{V}_{C_{i1}, D_i}^{(t)} \cdot \left[\hat{V}_i^{(t)}\right]^{-1} \cdot \left[D_i - \hat{\mu}_i^{(t)}\right] \tag{7.5.28}$$

$$\tilde{\sigma}_{C_1}^{2(t)} \triangleq VAR\left(C_{i1} \middle| D_i, \hat{\Theta}^{(t)}\right) = \hat{\sigma}_{R_1}^{2(t)} + \hat{V}_{C_{i1}, D_i}^{(t)} \cdot \left[\hat{V}_i^{(t)}\right]^{-1} \cdot \left[\hat{V}_{C_{i1}, D_i}^{(t)}\right]^T \tag{7.5.29}$$

The objective of M step is to find $\hat{\Theta}^{(t+1)}$ that maximizes $E\left[l_c | D_i, \hat{\Theta}^{(t)}\right]$, therefore

$$\hat{\sigma}_{\delta_i}^{2(t+1)} = \frac{\Sigma_{j=1}^{J}\Sigma_{k=1}^{K}\left[\left(\tilde{\delta}_{ij}^{k(t)}\right)^2 + \tilde{\sigma}_{\delta_{ij}^k}^{2(t)}\right]}{JK} \tag{7.5.30}$$

$$\hat{\sigma}_{\varepsilon_i}^{2(t+1)} = \frac{\Sigma_{j=1}^{J}\Sigma_{k=1}^{K}\left[\left(\tilde{\varepsilon}_{ij}^{k(t)}\right)^2 + \tilde{\sigma}_{\varepsilon_{ij}^k}^{2(t)}\right]}{JK} \tag{7.5.31}$$

$$\hat{\sigma}_{\tau_i}^{2(t+1)} = \frac{\Sigma_{j=1}^{J}\Sigma_{k=1}^{K}\left[\left(\tilde{\tau}_{ij}^{k(t)}\right)^2 + \tilde{\sigma}_{\tau_{ij}^k}^{2(t)}\right]}{JK} \tag{7.5.32}$$

$$\hat{\xi}_i^{(t+1)} = \frac{\Sigma_{j=1}^{J}\tilde{\xi}_{ij}^{(t)}}{J} \text{ and } \hat{\sigma}_{\xi_i}^{2(t+1)} = \frac{\Sigma_{j=1}^{J}\left[\left(\tilde{\xi}_{ij}^{(t)} - \hat{\xi}_i^{(t+1)}\right)^2 + \tilde{\sigma}_{\xi_i}^{2(t)}\right]}{J} \tag{7.5.33}$$

$$\hat{\beta}_1^{(t+1)} = \frac{\Sigma_{i=1}^{I}\tilde{B}_{i1}^{(t)}}{I} \text{ and } \hat{\gamma}_1^{(t+1)} = \frac{\Sigma_{i=1}^{I}\tilde{C}_{i1}^{(t)}}{I} \tag{7.5.34}$$

$$\hat{\sigma}_{R_1}^{2(t+1)} = \frac{\Sigma_{i=1}^{I}\left[\left(\tilde{A}_{i1}^{(t)} - 1\right)^2 + \tilde{\sigma}_{A_1}^{2(t)} + \left(\tilde{B}_{i1}^{(t)} - \hat{\beta}_1^{(t+1)}\right)^2 + \tilde{\sigma}_{B_1}^{2(t)} + \left(\tilde{C}_{i1}^{(t)} - \hat{\gamma}_1^{(t+1)}\right)^2 + \tilde{\sigma}_{C_1}^{2(t)}\right]}{3I} \tag{7.5.35}$$

where $\tilde{\delta}_{ij}^{k(t)}$ denotes $E\left[\delta_{ij}^k | D_i, \hat{\theta}^{(t)}\right]$ which could be obtained from (7.5.16), $\tilde{\sigma}_{\delta_{ij}^k}^{2(t)}$ denotes $VAR(\delta_{ij}^k | D_i, \hat{\theta}^{(t)})$ which could be obtained from (7.5.17), and similarly for $\tilde{\varepsilon}_{ij}^{k(t)}$, $\tilde{\sigma}_{\varepsilon_{ij}^k}^{2(t)}$, $\tilde{\tau}_{ij}^{k(t)}$, $\tilde{\sigma}_{\tau_{ij}^k}^{2(t)}$, $\tilde{\xi}_{ij}^{(t)}$ and $\tilde{\sigma}_{\xi_{ij}}^{2(t)}$.

It is worth noticing that α_0, β_0 and γ_0 did not appear in the likelihood function of completed data defined by (7.5.3), then in order to update their value, at the end of M step of each iteration, they would be replaced by solving likelihood function of observed data defined by (7.4.11). To be more specific, if denoting $V_i^{-1} \triangleq \Gamma_i = \begin{bmatrix} \Gamma_{i11} & \Gamma_{i12} & \Gamma_{i13} \\ \Gamma_{i21} & \Gamma_{i22} & \Gamma_{i23} \\ \Gamma_{i31} & \Gamma_{i32} & \Gamma_{i33} \end{bmatrix}$ with each block of

dimension $JK \times JK$, defining $S(M)$ as the function returning summation of all the elements in matrix M, and defining $ColS(M)$ as the function returning the column summation of all the columns in matrix M, then it could be obtained that

$$\begin{bmatrix} \hat{\alpha}_0^{(t+1)} \\ \hat{\beta}_0^{(t+1)} \\ \hat{\gamma}_0^{(t+1)} \end{bmatrix} = \left(\Sigma_{i=1}^{I} \begin{bmatrix} S\left(\hat{\Gamma}_{i11}^{(t)}\right) & S\left(\hat{\Gamma}_{i12}^{(t)}\right) & S\left(\hat{\Gamma}_{i13}^{(t)}\right) \\ S\left(\hat{\Gamma}_{i21}^{(t)}\right) & S\left(\hat{\Gamma}_{i22}^{(t)}\right) & S\left(\hat{\Gamma}_{i23}^{(t)}\right) \\ S\left(\hat{\Gamma}_{i31}^{(t)}\right) & S\left(\hat{\Gamma}_{i32}^{(t)}\right) & S\left(\hat{\Gamma}_{i33}^{(t)}\right) \end{bmatrix} \right)^{-1} \cdot \Sigma_{i=1}^{I} \begin{bmatrix} ColS\left[\hat{\Gamma}_{i11}^{(t)}, \hat{\Gamma}_{i12}^{(t)}, \hat{\Gamma}_{i13}^{(t)}\right] \cdot \left(D_i - \tilde{\mu}_i^{(t)}\right) \\ ColS\left[\hat{\Gamma}_{i21}^{(t)}, \hat{\Gamma}_{i22}^{(t)}, \hat{\Gamma}_{i23}^{(t)}\right] \cdot \left(D_i - \hat{\beta}_1^{(t+1)} \tilde{\mu}_i^{(t)}\right) \\ ColS\left[\hat{\Gamma}_{i31}^{(t)}, \hat{\Gamma}_{i32}^{(t)}, \hat{\Gamma}_{i33}^{(t)}\right] \cdot \left(D_i - \hat{\gamma}_1^{(t+1)} \tilde{\mu}_i^{(t)}\right) \end{bmatrix} \tag{7.5.36}$$

To prove (7.5.36), from (7.4.11), it could be obtained that

$$\frac{\partial l}{\partial \alpha_0} = \Sigma_{i=1}^{I} \left(\left[1_{JK}^{T}, 0_{1\times JK}, 0_{1\times JK} \right] \cdot \Gamma_i \cdot [D_i - \mu_i] \right)$$

$$= \Sigma_{i=1}^{I} \left(ColS[\Gamma_{i11}, \Gamma_{i12}, \Gamma_{i13}] \cdot \left(D_i - \begin{bmatrix} \xi_i 1_{JK} \\ \beta_1 \xi_i 1_{JK} \\ \gamma_1 \xi_i 1_{JK} \end{bmatrix} - \begin{bmatrix} \alpha_0 1_{JK} \\ \beta_0 1_{JK} \\ \gamma_0 1_{JK} \end{bmatrix} \right) \right)$$

and setting it to zero will yield

$$\Sigma_{i=1}^{I} [S(\Gamma_{i11}), S(\Gamma_{i12}), S(\Gamma_{i13})] \cdot \begin{bmatrix} \alpha_0 \\ \beta_0 \\ \gamma_0 \end{bmatrix} = \Sigma_{i=1}^{I} \left(ColS[\Gamma_{i11}, \Gamma_{i12}, \Gamma_{i13}] \cdot \left(D_i - \begin{bmatrix} \xi_i 1_{JK} \\ \beta_1 \xi_i 1_{JK} \\ \gamma_1 \xi_i 1_{JK} \end{bmatrix} \right) \right)$$

Similarly, from $\frac{\partial l}{\partial \beta_0} = 0$ and $\frac{\partial l}{\partial \gamma_0} = 0$, we have

$$\Sigma_{i=1}^{I} [S(\Gamma_{i21}), S(\Gamma_{i22}), S(\Gamma_{i23})] \cdot \begin{bmatrix} \alpha_0 \\ \beta_0 \\ \gamma_0 \end{bmatrix} = \Sigma_{i=1}^{I} \left(ColS[\Gamma_{i21}, \Gamma_{i22}, \Gamma_{i23}] \cdot \left(D_i - \begin{bmatrix} \xi_i 1_{JK} \\ \beta_1 \xi_i 1_{JK} \\ \gamma_1 \xi_i 1_{JK} \end{bmatrix} \right) \right)$$

and

$$\Sigma_{i=1}^{I}[S(\Gamma_{i31}), S(\Gamma_{i32}), S(\Gamma_{i33})] \cdot \begin{bmatrix}\alpha_0\\ \beta_0\\ \gamma_0\end{bmatrix} = \Sigma_{i=1}^{I}\left(ColS[\Gamma_{i31}, \Gamma_{i32}, \Gamma_{i33}] \cdot \left(D_i - \begin{bmatrix}\xi_i 1_{JK}\\ \beta_1\xi_i 1_{JK}\\ \gamma_1\xi_i 1_{JK}\end{bmatrix}\right)\right)$$

As a result, it could be shown that

$$\Sigma_{i=1}^{I}\begin{bmatrix}S(\Gamma_{i11}) & S(\Gamma_{i12}) & S(\Gamma_{i13})\\ S(\Gamma_{i21}) & S(\Gamma_{i22}) & S(\Gamma_{i23})\\ S(\Gamma_{i31}) & S(\Gamma_{i32}) & S(\Gamma_{i33})\end{bmatrix} \cdot \begin{bmatrix}\alpha_0\\ \beta_0\\ \gamma_0\end{bmatrix} = \Sigma_{i=1}^{I}\left(\begin{bmatrix} ColS[\Gamma_{i11}, \Gamma_{i12}, \Gamma_{i13}]\\ ColS[\Gamma_{i21}, \Gamma_{i22}, \Gamma_{i23}]\\ ColS[\Gamma_{i31}, \Gamma_{i32}, \Gamma_{i33}]\end{bmatrix} \cdot \left(D_i - \begin{bmatrix}\xi_i 1_{JK}\\ \beta_1\xi_i 1_{JK}\\ \gamma_1\xi_i 1_{JK}\end{bmatrix}\right)\right)$$

which subsequently yields (7.5.36).

When the difference between estimates of two consecutive steps, i.e. $\widehat{\Theta}^{(t)}$ and $\widehat{\Theta}^{(t+1)}$ is smaller than a certain tolerance, 1e-8 in this study, EM algorithm reaches convergence.

Upon convergence, prediction of elements in Λ defined by (7.5.2) is a by-product of EM algorithm, where A_{i1}, B_{i1} and C_{i1} are of major interest since they imply the relation between measurements and true abundance of each individual bacterial across all three platforms. From (7.5.24), (7.5.26) and (7.5.28), it is obvious that

$$\tilde{A}_{i1}^{(N)} \triangleq E\left[A_{i1}|D_i, \hat{\Theta}^{(N)}\right] = 1 + \hat{V}_{A_{i1},D_i}^{(N)} \cdot \left[\hat{V}_i^{(N)}\right]^{-1} \cdot \left[D_i - \hat{\mu}_i^{(N)}\right] \tag{7.5.37}$$

$$\tilde{B}_{i1}^{(N)} \triangleq E\left[B_{i1}|D_i, \hat{\Theta}^{(N)}\right] = \hat{\beta}_1^{(N)} + \hat{V}_{B_{i1},D_i}^{(N)} \cdot \left[\hat{V}_i^{(N)}\right]^{-1} \cdot \left[D_i - \hat{\mu}_i^{(N)}\right] \tag{7.5.38}$$

$$\tilde{C}_{i1}^{(N)} \triangleq E\left[C_{i1}|D_i, \hat{\Theta}^{(N)}\right] = \hat{\gamma}_1^{(N)} + \hat{V}_{C_{i1},D_i}^{(N)} \cdot \left[\hat{V}_i^{(N)}\right]^{-1} \cdot \left[D_i - \hat{\mu}_i^{(N)}\right] \tag{7.5.39}$$

where N is the number of steps for EM algorithm to converge.

7.6 Hypothesis test

After obtaining estimates, it is of our interest know whether V1V2, V1V3 and V3V4 are consistent or not, and thus there are four related hypothesis tests, which are

(1) $H_0: (\alpha_0, 1) = (\beta_0, \beta_1) = (\gamma_0, \gamma_1)$ is used to test whether V1V2, V1V3 and V3V4 are consistent together

(2) $H_0: (\alpha_0, 1) = (\beta_0, \beta_1)$ is used to test whether V1V2 and V1V3 are consistent;

(3) $H_0: (\alpha_0, 1) = (\gamma_0, \gamma_1)$ is used to test whether V1V2 and V3V4 are consistent;

(4) $H_0: (\beta_0, \beta_1) = (\gamma_0, \gamma_1)$ is used to test whether V1V3 and V3V4 are consistent.

Likelihood ratio test (LRT) is adopted to test each one of them [33], where $-2(l_0 - l_1) \dot{\sim} \chi^2_{df}$ with l_0 being the log likelihood under the null hypothesis, l_1 being the log likelihood without any restriction, and df being the degrees of freedom lost when applying the restrictions in H_0.

Under the null hypothesis of (1), to obtain the corresponding estimates, (7.5.34) should be modified to $\hat{\beta}_1^{(t+1)} = \hat{\gamma}_1^{(t+1)} = 1$. As for α_0, β_0 and γ_0, since under H_0 they are identical, then similar to the process of deriving (7.5.36), in (7.4.11) if defining $\alpha_0 = \beta_0 = \gamma_0 = \Delta_0$, it follows that

$$\frac{\partial l}{\partial \Delta_0} = \Sigma_{i=1}^{I}\left(\left[1_{JK}^T, 1_{JK}^T, 1_{JK}^T\right] \cdot \Gamma_i \cdot [D_i - \mu_i]\right) = \Sigma_{i=1}^{I}(ColS(\Gamma_i) \cdot [D_i - \mu_i]) \qquad (7.6.1)$$

thus by $\frac{\partial l}{\partial \Delta_0} = 0$ we have

$$\Sigma_{i=1}^{N} S(\Gamma_i) \cdot \Delta_0 = \Sigma_{i=1}^{I}\left(ColS(\Gamma_i) \cdot \left(D_i - \begin{bmatrix} \xi_i 1_{JK} \\ \beta_1 \xi_i 1_{JK} \\ \gamma_1 \xi_i 1_{JK} \end{bmatrix}\right)\right) \qquad (7.6.2)$$

As a result, (7.6.36) should be modified to

$$\hat{\alpha}_0^{(t+1)} = \hat{\beta}_0^{(t+1)} = \hat{\gamma}_0^{(t+1)} = \frac{\Sigma_{i=1}^{I}\left(ColS(\Gamma_i) \cdot \left(D_i - \begin{bmatrix} \xi_i 1_{JK} \\ \beta_1 \xi_i 1_{JK} \\ \gamma_1 \xi_i 1_{JK} \end{bmatrix}\right)\right)}{\Sigma_{i=1}^{N} S(\Gamma_i)} \qquad (7.6.3)$$

Due to H_0: $(\alpha_0, 1) = (\beta_0, \beta_1) = (\gamma_0, \gamma_1)$, the degrees of freedom lost is 4, i.e. $df = 4$.

As for the null hypothesis of (2), $\hat{\beta}_1^{(t+1)}$ in (7.5.34) should always be kept at 1, and if defining $\alpha_0 = \beta_0 = \Delta_0$, then similarly from $\frac{\partial l}{\partial \Delta_0} = \Sigma_{i=1}^{I}\left(\left[1_{JK}^T, 1_{JK}^T, 0_{1\times JK}\right] \cdot \Gamma_i \cdot [D_i - \mu_i]\right) = 0$ it could be obtained that

$$\Sigma_{i=1}^{I}\left[S\left(\begin{bmatrix} \Gamma_{i11} & \Gamma_{i12} \\ \Gamma_{i21} & \Gamma_{i22} \end{bmatrix}\right), S\left(\begin{bmatrix} \Gamma_{i13} \\ \Gamma_{i23} \end{bmatrix}\right)\right] \cdot \begin{bmatrix} \Delta_0 \\ \gamma_0 \end{bmatrix}$$

$$= \Sigma_{i=1}^{I}\left(ColS\left(\begin{bmatrix} \Gamma_{i11} & \Gamma_{i12} & \Gamma_{i13} \\ \Gamma_{i21} & \Gamma_{i22} & \Gamma_{i23} \end{bmatrix}\right) \cdot \left(D_i - \begin{bmatrix} \xi_i 1_{JK} \\ \beta_1 \xi_i 1_{JK} \\ \gamma_1 \xi_i 1_{JK} \end{bmatrix}\right)\right)$$

And from $\frac{\partial l}{\partial \gamma_0} = \Sigma_{i=1}^{I}\left(\left[0_{1\times JK}, 0_{1\times JK}, 1_{JK}^T\right] \cdot \Gamma_i \cdot [D_i - \mu_i]\right) = 0$ we have

$$\Sigma_{i=1}^{I}[S([\Gamma_{i31} \quad \Gamma_{i32}]), S(\Gamma_{i33})] \cdot \begin{bmatrix}\Delta_0 \\ \gamma_0\end{bmatrix} = \Sigma_{i=1}^{I}\left(ColS([\Gamma_{i31} \quad \Gamma_{i32} \quad \Gamma_{i33}]) \cdot \left(D_i - \begin{bmatrix}\xi_i 1_{JK} \\ \beta_1 \xi_i 1_{JK} \\ \gamma_1 \xi_i 1_{JK}\end{bmatrix}\right)\right)$$

Together it could generate that

$$\Sigma_{i=1}^{I}\begin{bmatrix} S\left(\begin{bmatrix}\Gamma_{i11} & \Gamma_{i12} \\ \Gamma_{i21} & \Gamma_{i22}\end{bmatrix}\right) & S\left(\begin{bmatrix}\Gamma_{i13} \\ \Gamma_{i23}\end{bmatrix}\right) \\ S([\Gamma_{i31} \quad \Gamma_{i32}]) & S(\Gamma_{i33}) \end{bmatrix} \cdot \begin{bmatrix}\Delta_0 \\ \gamma_0\end{bmatrix}$$

$$= \Sigma_{i=1}^{I}\left(\begin{bmatrix} ColS\left(\begin{bmatrix}\Gamma_{i11} & \Gamma_{i12} & \Gamma_{i13} \\ \Gamma_{i21} & \Gamma_{i22} & \Gamma_{i23}\end{bmatrix}\right) \\ ColS([\Gamma_{i31} \quad \Gamma_{i32} \quad \Gamma_{i33}]) \end{bmatrix} \cdot \left(D_i - \begin{bmatrix}\xi_i 1_{JK} \\ \beta_1 \xi_i 1_{JK} \\ \gamma_1 \xi_i 1_{JK}\end{bmatrix}\right)\right)$$

Therefore, (7.6.36) should be updated to

$$\begin{bmatrix}\widehat{\Delta}_0^{(t+1)} \\ \hat{\gamma}_0^{(t+1)}\end{bmatrix} = \Sigma_{i=1}^{I}\left(\begin{bmatrix} ColS\left(\begin{bmatrix}\hat{\Gamma}_{i11}^{(t)} & \hat{\Gamma}_{i12}^{(t)} & \hat{\Gamma}_{i13}^{(t)} \\ \hat{\Gamma}_{i21}^{(t)} & \hat{\Gamma}_{i22}^{(t)} & \hat{\Gamma}_{i23}^{(t)}\end{bmatrix}\right) \\ ColS\left([\hat{\Gamma}_{i31}^{(t)} \quad \hat{\Gamma}_{i32}^{(t)} \quad \hat{\Gamma}_{i33}^{(t)}]\right) \end{bmatrix} \cdot \left(D_i - \begin{bmatrix}\hat{\xi}_i^{(t)} 1_{JK} \\ \hat{\beta}_1^{(t)} \hat{\xi}_\iota^{(t)} 1_{JK} \\ \hat{\gamma}_1^{(t)} \hat{\xi}_i^{(t)} 1_{JK}\end{bmatrix}\right)\right)$$

$$\cdot \left(\Sigma_{i=1}^{I}\begin{bmatrix} S\left(\begin{bmatrix}\hat{\Gamma}_{i11}^{(t)} & \hat{\Gamma}_{i12}^{(t)} \\ \hat{\Gamma}_{i21}^{(t)} & \hat{\Gamma}_{i22}^{(t)}\end{bmatrix}\right) & S\left(\begin{bmatrix}\hat{\Gamma}_{i13}^{(t)} \\ \hat{\Gamma}_{i23}^{(t)}\end{bmatrix}\right) \\ S\left([\hat{\Gamma}_{i31}^{(t)} \quad \hat{\Gamma}_{i32}^{(t)}]\right) & S\left(\hat{\Gamma}_{i33}^{(t)}\right) \end{bmatrix}\right)^{-1}$$

and $\hat{\alpha}^{(t+1)} = \hat{\beta}^{(t+1)} = \widehat{\Delta}_0^{(t+1)}$ with $df = 2$.

In parallel, for the test in (3), $\hat{\gamma}_1^{(t+1)}$ should always be 1, (7.6.36) should be changed to

$$\begin{bmatrix} \widehat{\Delta}_0^{(t+1)} \\ \hat{\beta}_0^{(t+1)} \end{bmatrix} = \Sigma_{i=1}^{I} \left(\begin{bmatrix} ColS\left(\begin{bmatrix} \hat{\Gamma}_{i11}^{(t)} & \hat{\Gamma}_{i12}^{(t)} & \hat{\Gamma}_{i13}^{(t)} \\ \hat{\Gamma}_{i31}^{(t)} & \hat{\Gamma}_{i32}^{(t)} & \hat{\Gamma}_{i33}^{(t)} \end{bmatrix} \right) \\ ColS\left(\begin{bmatrix} \hat{\Gamma}_{i21}^{(t)} & \hat{\Gamma}_{i22}^{(t)} & \hat{\Gamma}_{i23}^{(t)} \end{bmatrix} \right) \end{bmatrix} \cdot \left(D_i - \begin{bmatrix} \hat{\xi}_i^{(t)} 1_{JK} \\ \hat{\beta}_1^{(t)} \hat{\xi}_\iota^{(t)} 1_{JK} \\ \hat{\gamma}_1^{(t)} \hat{\xi}_i^{(t)} 1_{JK} \end{bmatrix} \right) \right)$$

$$\cdot \left(\Sigma_{i=1}^{I} \begin{bmatrix} S\left(\begin{bmatrix} \hat{\Gamma}_{i11}^{(t)} & \hat{\Gamma}_{i13}^{(t)} \\ \hat{\Gamma}_{i31}^{(t)} & \hat{\Gamma}_{i33}^{(t)} \end{bmatrix} \right) & S\left(\begin{bmatrix} \hat{\Gamma}_{i12}^{(t)} \\ \hat{\Gamma}_{i32}^{(t)} \end{bmatrix} \right) \\ S\left(\begin{bmatrix} \hat{\Gamma}_{i21}^{(t)} & \hat{\Gamma}_{i23}^{(t)} \end{bmatrix} \right) & S\left(\hat{\Gamma}_{i22}^{(t)} \right) \end{bmatrix} \right)^{-1}$$

and $\hat{\alpha}_0^{(t+1)} = \hat{\gamma}_0^{(t+1)} = \widehat{\Delta}_0^{(t+1)}$ with $df = 2$. As for test in (4), (7.5.34) should be modified to

$$\hat{\beta}_1^{(t+1)} = \hat{\gamma}_1^{(t+1)} = \frac{\Sigma_{i=1}^{I} \left(\tilde{B}_{i1}^{(t)} + \tilde{C}_{i1}^{(t)} \right)}{2I}$$

And (7.6.36) should be updated to

$$\begin{bmatrix} \hat{\alpha}_0^{(t+1)} \\ \widehat{\Delta}_0^{(t+1)} \end{bmatrix} = \Sigma_{i=1}^{I} \left(\begin{bmatrix} ColS\left(\begin{bmatrix} \hat{\Gamma}_{i11}^{(t)} & \hat{\Gamma}_{i12}^{(t)} & \hat{\Gamma}_{i13}^{(t)} \end{bmatrix} \right) \\ ColS\left(\begin{bmatrix} \hat{\Gamma}_{i21}^{(t)} & \hat{\Gamma}_{i22}^{(t)} & \hat{\Gamma}_{i23}^{(t)} \\ \hat{\Gamma}_{i31}^{(t)} & \hat{\Gamma}_{i32}^{(t)} & \hat{\Gamma}_{i33}^{(t)} \end{bmatrix} \right) \end{bmatrix} \cdot \left(D_i - \begin{bmatrix} \hat{\xi}_i^{(t)} 1_{JK} \\ \hat{\beta}_1^{(t)} \hat{\xi}_\iota^{(t)} 1_{JK} \\ \hat{\gamma}_1^{(t)} \hat{\xi}_i^{(t)} 1_{JK} \end{bmatrix} \right) \right)$$

$$\cdot \left(\Sigma_{i=1}^{I} \begin{bmatrix} S\left(\hat{\Gamma}_{i11}^{(t)} \right) & S\left(\begin{bmatrix} \hat{\Gamma}_{i12}^{(t)} & \hat{\Gamma}_{i13}^{(t)} \end{bmatrix} \right) \\ S\left(\begin{bmatrix} \hat{\Gamma}_{i21}^{(t)} \\ \hat{\Gamma}_{i31}^{(t)} \end{bmatrix} \right) & S\left(\begin{bmatrix} \hat{\Gamma}_{i22}^{(t)} & \hat{\Gamma}_{i23}^{(t)} \\ \hat{\Gamma}_{i32}^{(t)} & \hat{\Gamma}_{i33}^{(t)} \end{bmatrix} \right) \end{bmatrix} \right)^{-1}$$

and $\hat{\beta}_0^{(t+1)} = \hat{\gamma}_0^{(t+1)} = \widehat{\Delta}_0^{(t+1)}$ with $df = 2$.

7.7 Generalized method of moments

To simplify the model, sample means of all the technical replicates are computed and thus (7.4.1) could be expressed as

$$\begin{cases} \bar{X}_{ij} = \alpha_0 + A_{i1}\xi_{ij} + \bar{\delta}_{ij} \\ \bar{Y}_{ij} = \beta_0 + B_{i1}\xi_{ij} + \bar{\varepsilon}_{ij} \\ \bar{Z}_{ij} = \gamma_0 + C_{i1}\xi_{ij} + \bar{\tau}_{ij} \end{cases} \tag{7.7.1}$$

To establish orthogonal conditions for (7.7.1), (A_{i1}, B_{i1}, C_{i1}), or equivalently (a_{i1}, b_{i1}, c_{i1}) are treated as fixed parameters instead of random variables here and (7.7.1) is re-arranged as

$$\begin{cases} \bar{Y}_{ij} = \beta_{i0}^* + \beta_{i1}^* \bar{X}_{ij} + \varepsilon_{ij}^* \\ \bar{Z}_{ij} = \gamma_{i0}^* + \gamma_{i1}^* \bar{X}_{ij} + \tau_{ij}^* \end{cases} \tag{7.7.2}$$

where

$$\beta_{i0}^* = \beta_0 - \frac{\beta_1 + b_{i1}}{1 + a_{i1}}\alpha_0 \text{ and } \gamma_{i0}^* = \gamma_0 - \frac{\gamma_1 + c_{i1}}{1 + a_{i1}}\alpha_0 \tag{7.7.3}$$

$$\beta_{i1}^* = \frac{\beta_1 + b_{i1}}{1 + a_{i1}} \text{ and } \gamma_{i0}^* = \frac{\gamma_1 + c_{i1}}{1 + a_{i1}} \tag{7.7.4}$$

$$\bar{\varepsilon}_{ij}^* = \bar{\varepsilon}_{ij} - \frac{\beta_1 + b_{i1}}{1 + a_{i1}}\delta_{ij} \text{ and } \bar{\tau}_{ij}^* = \bar{\tau}_{ij} - \frac{\gamma_1 + c_{i1}}{1 + a_{i1}}\bar{\delta}_{ij} \tag{7.7.5}$$

It is not hard to see that $cov\left(\bar{Z}_{ij}, \bar{\varepsilon}_{ij}^*\right) = cov\left(\bar{Y}_{ij}, \bar{\tau}_{ij}^*\right) = 0$, thus $\bar{Z}_{ij}$ and $\bar{Y}_{ij}$ could be considered as the instrumental variables for the two equations in (7.7.2) respectively, together with $E\left[\bar{Y}_{ij}\right] = \beta_{i0}^* + \beta_{i1}^* E\left[\bar{X}_{ij}\right]$ and $E\left[\bar{Z}_{ij}\right] = \gamma_{i0}^* + \gamma_{i1}^* E\left[\bar{X}_{ij}\right]$, we have

$$E\left[\bar{Y}_{ij}\right] - \beta_{i0}^* - \beta_{i1}^* E\left[\bar{X}_{ij}\right] = 0 \tag{7.7.6}$$

$$E\left[\bar{Z}_{ij}\right] - \gamma_{i0}^{*} - \gamma_{i1}^{*} E\left[\bar{X}_{ij}\right] = 0 \tag{7.7.7}$$

$$E\left[\bar{Z}_{ij} \cdot \left(\bar{Y}_{ij} - \beta_{i0}^{*} - \beta_{i1}^{*} \bar{X}_{ij}\right)\right] = 0 \tag{7.7.8}$$

$$E\left[\bar{Y}_{ij} \cdot \left(\bar{Z}_{ij} - \gamma_{i0}^{*} - \gamma_{i1}^{*} \bar{X}_{ij}\right)\right] = 0 \tag{7.7.9}$$

where substituting the population expectations with sample ones will give

$$\bar{\bar{Y}}_i - \beta_{i0}^{*} - \beta_{i1}^{*} \bar{\bar{X}}_i = 0 \tag{7.7.10}$$

$$\bar{\bar{Z}}_i - \gamma_{i0}^{*} - \gamma_{i1}^{*} \bar{\bar{X}}_i = 0 \tag{7.7.11}$$

$$\overline{\overline{YZ}}_i - \beta_{i0}^{*} \bar{\bar{Z}}_i - \beta_{i1}^{*} \overline{\overline{XZ}}_i = 0 \tag{7.7.12}$$

$$\overline{\overline{YZ}}_i - \gamma_{i0}^{*} \bar{\bar{Y}}_i - \gamma_{i1}^{*} \overline{\overline{XY}}_i = 0 \tag{7.7.13}$$

where $\bar{\bar{X}}_i = \frac{\Sigma_{j=1}^{J} \bar{X}_{ij}}{J}$, $\bar{\bar{Y}}_i = \frac{\Sigma_{j=1}^{J} \bar{Y}_{ij}}{J}$, $\bar{\bar{Z}}_i = \frac{\Sigma_{j=1}^{J} \bar{Z}_{ij}}{J}$, $\overline{\overline{XY}}_i = \frac{\Sigma_{j=1}^{J} \bar{X}_{ij} \bar{Y}_{ij}}{J}$, $\overline{\overline{XZ}}_i = \frac{\Sigma_{j=1}^{J} \bar{X}_{ij} \bar{Z}_{ij}}{J}$ and $\overline{\overline{YZ}}_i = \frac{\Sigma_{j=1}^{J} \bar{Y}_{ij} \bar{Z}_{ij}}{J}$. Apparently (7.7.10) – (7.6.13) could not be satisfied simultaneously for $i = 1, \cdots, I$,

therefore the GMM estimator $\widehat{\Theta}_{GMM}$ should be the one which minimizes the weighted sum of squares of them.

Following the same strategy used in two step efficient GMM provided in Section 5.4.1, an intuitive initial estimate of Θ would be of no weights, i.e.

$$\widehat{\Theta}_{init} = \underset{\Theta}{argmin}\, \Sigma_{i=1}^{I} [(\bar{\bar{Y}}_i - \beta_{i0}^{*} - \beta_{i1}^{*} \bar{\bar{X}}_i)^2 + (\bar{\bar{Z}}_i - \gamma_{i0}^{*} - \gamma_{i1}^{*} \bar{\bar{X}}_i)^2 + (\overline{\overline{YZ}}_i - \beta_{i0}^{*} \bar{\bar{Z}}_i - \beta_{i1}^{*} \overline{\overline{XZ}}_i)^2 +$$

$$(\overline{\overline{YZ}}_i - \gamma_{i0}^{*} \bar{\bar{Y}}_i - \gamma_{i1}^{*} \overline{\overline{XY}}_i)^2] \tag{7.7.14}$$

Because (7.7.10) and (7.7.12) are generated based on (7.7.6) and (7.7.8) for $j = 1, \cdots, J$, then a natural weight matrix W_{i1} would be the inverse of the sample covariance matrix of left hand sides of (7.7.6) and (7.7.8) for $j = 1, \cdots, J$ after substituting Θ with $\widehat{\Theta}_{init}$, and similarly W_{i2} is available for (7.7.11) and (7.7.13). Finally, the GMM estimator $\widehat{\Theta}$ should be

$$\widehat{\Theta} = \underset{\Theta}{argmin}\, \Sigma_{i=1}^{I} [D_{i1}^T W_{i1} D_{i1} + D_{i2}^T W_{i2} D_{i2}] \tag{7.7.15}$$

where $D_{i1} = [\bar{Y}_i - \beta_{i0}^* - \beta_{i1}^* \bar{X}_i, \overline{YZ}_i - \beta_{i0}^* \bar{Z}_i - \beta_{i1}^* \overline{XZ}_i]^T$ and $D_{i2} = [\bar{Z}_i - \gamma_{i0}^* - \gamma_{i1}^* \bar{X}_i, \overline{YZ}_i - \gamma_{i0}^* \bar{Y}_i - \gamma_{i1}^* \overline{XY}_i]^T$.

To illustrate the quality of this method of moment estimator, data was simulated 20 times based on the MLE parameter estimates obtained by EM algorithm, and its comparison with MLE results in terms of α_0, β_0, γ_0, β_1 and γ_1, is shown in Figure 7.3A–E.

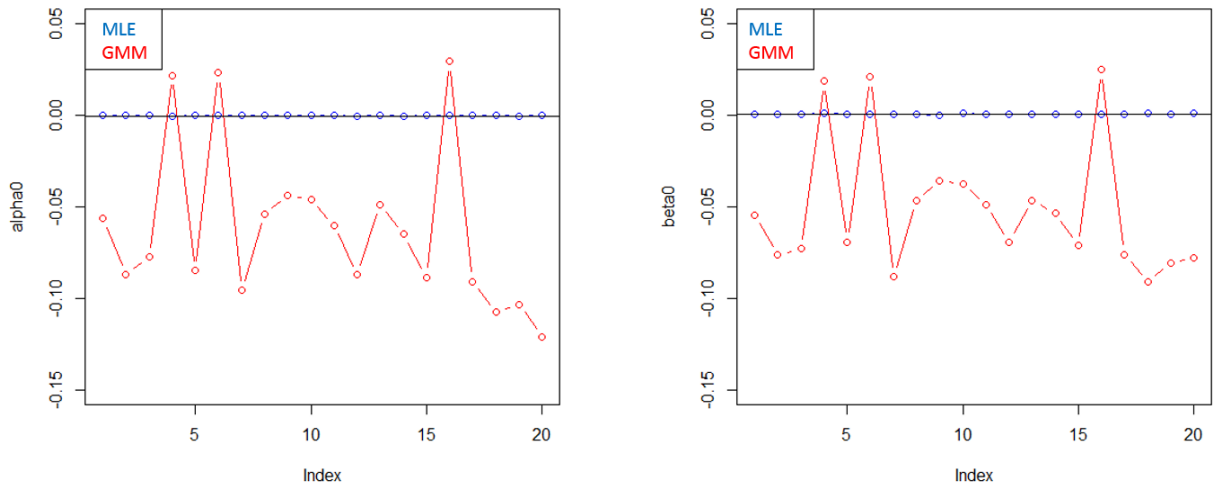

Figure 7.3. A (left) – Comparison of $\hat{\alpha}_0$ between MLE and GMM; B (right) – Comparison of $\hat{\beta}_0$ between MLE and GMM

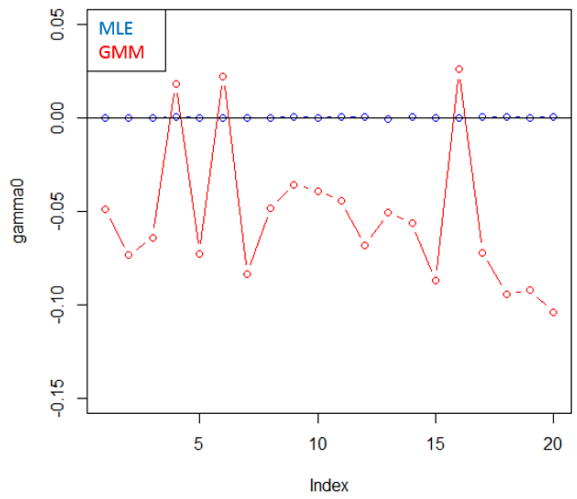

Figure 7.3C – Comparison of $\hat{\gamma}_0$ between MLE and GMM

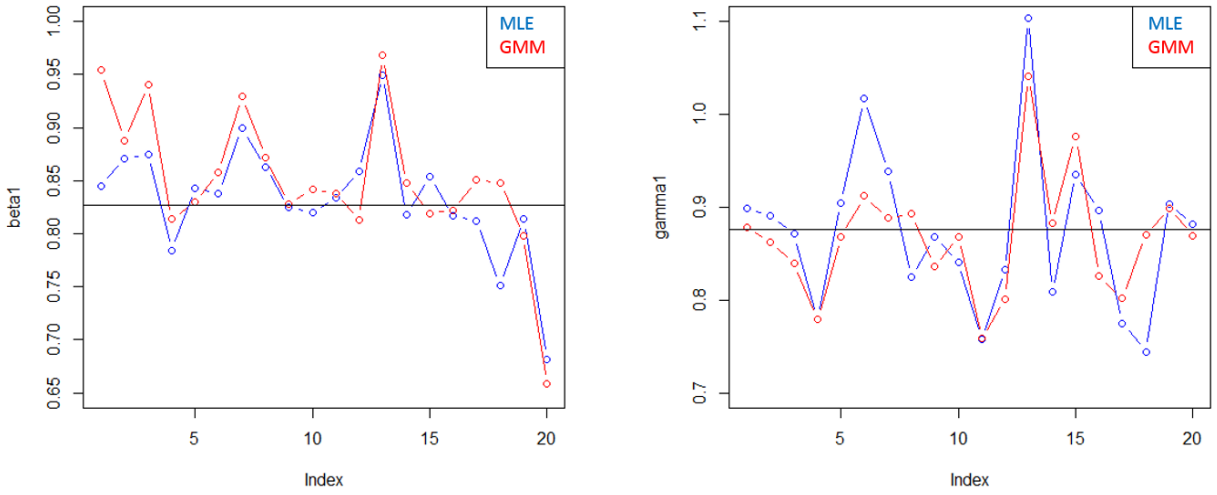

Figure 7.3. D (left) – Comparison of $\hat{\beta}_1$ between MLE and GMM; E (right) – Comparison of $\hat{\gamma}_1$ between MLE and GMM

It could be seen that GMM and MLE perform quite consistently across 20 simulations, and to further demonstrate the quality of GMM estimators, 500 simulations under the same setting but on GMM alone was run, and the MSE of α_0, β_0, γ_0, β_1 and γ_1 are $5.1e-3$, $3.7e-3$, $4.0e-3$, $6.3e-3$ and $5.7e-3$ respectively, indicating reasonable quality.

7.8 Pairwise Comparison

To perform the similar analysis when only two platforms are available, and to further test the outputs generated by the model above, EIV with random effects is proposed here to handle the problem. For example with data from V1V2 and V1V3, the model would be

$$\begin{cases} X_{ij}^k = \alpha_0 + A_{i1}\xi_{ij} + \delta_{ij}^k \\ Y_{ij}^k = \beta_0 + B_{i1}\xi_{ij} + \varepsilon_{ij}^k \end{cases} \tag{7.8.1}$$

with $A_{i1} = 1 + a_{i1}$ and $B_{i1} = \beta_1 + b_{i1}$ like in Section 7.4. EM algorithm in Section 7.5 could be applied for estimation, after which $H_0: (\alpha_0, 1) = (\beta_0, \beta_1)$ versus $H_1: (\alpha_0, 1) \neq (\beta_0, \beta_1)$ could used to test whether V1V2 and V1V3 are consistent. Similarly for V1V2 versus V3V4, and V1V3 versus V3V4.

7.9 Reliability

Conditioning on each A_{i1}, B_{i1} and C_{i1}, it follows naturally that the reliabilities of each platform for this particular bacterium i would be, $R_{X_i}^2 = \frac{A_{i1}^2\sigma_{\xi_i}^2}{A_{i1}^2\sigma_{\xi_i}^2+\sigma_{\delta_i}^2}$, $R_{Y_i}^2 = \frac{B_{i1}^2\sigma_{\xi_i}^2}{B_{i1}^2\sigma_{\xi_i}^2+\sigma_{\varepsilon_i}^2}$ and $R_{Z_i}^2 = \frac{C_{i1}^2\sigma_{\xi_i}^2}{C_{i1}^2\sigma_{\xi_i}^2+\sigma_{\tau_i}^2}$ for V1V2, V1V3 and V3V4 respectively. A naïve estimates on overall reliability of each platform across all bacteria would be $\frac{\Sigma_{i=1}^I R_{X_i}^2}{I}$, $\frac{\Sigma_{i=1}^I R_{Y_i}^2}{I}$ and $\frac{\Sigma_{i=1}^I R_{Z_i}^2}{I}$, but that does not take into account any weight on each i.

Since A_{i1}, B_{i1} and C_i are all random slopes with probability density function $f(A_{i1}) = \frac{1}{\sigma_{R_1}\sqrt{2\pi}} e^{-\frac{(A_{i1}-1)^2}{2\sigma_{R_1}^2}}$, $f(B_{i1}) = \frac{1}{\sigma_{R_1}\sqrt{2\pi}} e^{-\frac{(B_{i1}-\beta_1)^2}{2\sigma_{R_1}^2}}$ and $f(C_{i1}) = \frac{1}{\sigma_{R_1}\sqrt{2\pi}} e^{-\frac{(C_{i1}-\gamma_1)^2}{2\sigma_{R_1}^2}}$, then these pdf's could be used as weights for each i. Consequently, the overall reliability of each platform is defined as $R_X^2 = \frac{\Sigma_{i=1}^I R_{X_i}^2 f(A_{i1})}{\Sigma_{i=1}^I f(A_{i1})}$, $R_Y^2 = \frac{\Sigma_{i=1}^I R_{Y_i}^2 f(B_{i1})}{\Sigma_{i=1}^I f(B_{i1})}$ and $R_Z^2 = \frac{\Sigma_{i=1}^I R_{Z_i}^2 f(C_{i1})}{\Sigma_{i=1}^I f(C_{i1})}$, where denominators are used to guarantee the range of reliability is from 0 to 1.

7.10 Results

Upon completion of the EM algorithm proposed in Section 7.5, it could be obtained that $\hat{\alpha}_0 = -3.5e-4$, $\hat{\beta}_0 = 5.8e-4$, $\hat{\gamma}_0 = 2.2e-4$, $\hat{\beta}_1 = 0.83$ and $\hat{\gamma}_1 = 0.88$, meaning that compared with V1V2, both V1V3 and V3V4 tended to underestimate the abundance level in average.

Test (1) in Section 7.6 generated p value of $1.03e-5$, meaning V1V2, V1V3 and V3V4 do not have overall consistency, while test (2), (3) and (4) gave p values of $2.4e-7$, 0.105, and 0.063 respectively, indicating the consistency between V1V2 and V3V4, V1V3 and V3V4, but the discrepancy between V1V2 and V1V3.

Figure 7.4 shows the relation between estimated mean of abundance of all the bacteria, i.e. $\hat{\xi}_i, i = 1, \cdots, I$, and the corresponding predicted slopes for all three platforms, i.e.

$A_{i1}, B_{i1}, C_{i1}, i = 1, \cdots, I$ obtained by (7.5.37) – (7.5.39), and it indicates the existence of significant proportional systematic errors when the abundance is low and hence the necessity to include random slopes for each bacterium.

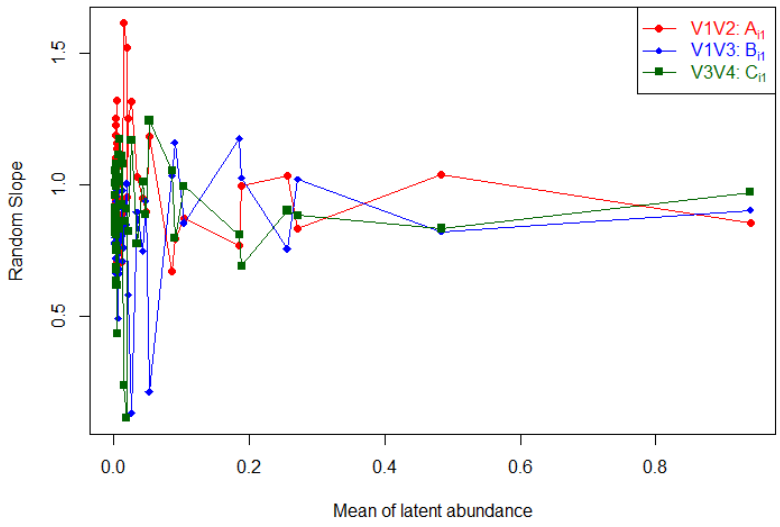

Figure 7.4. Scatter plot of estimated $\xi_i's$ versus A_{i1}, B_{i1}, C_{i1}.

Table 7.2 includes the method of moments estimates and their bootstrap confidence intervals [15] obtained by procedures described in Section 7.7. Unfortunately the corresponding hypothesis tests like in Section 7.6 is still undeveloped, but the estimations did show the same pattern as the ones from EM algorithm, that is compared with V1V2, V1V3 and V3V4 tended to underestimate the abundance.

Table 7.2. Method of moments estimates and the corresponding bootstrap confidence intervals

	Estimates	Bootstrap confidence interval
$\hat{\alpha}_0$	$-3.16e-4$	(-0.0038, 0.0101)
$\hat{\beta}_0$	$-2.96e-4$	(-0.0032, 0.0087)
$\hat{\gamma}_0$	$-2.88e-4$	(-0.0035, 0.0094)
$\hat{\beta}_1$	0.94	(0.6719, 1.1305)
$\hat{\gamma}_1$	0.92	(0.7950, 1.0932)

Pairwise comparison with two platforms analyzed at a time as mentioned in Section 7.8 generates coherent results. As Table 7.3 shows, V3V4 is consistent with V1V2 and V1V3, while V1V2 and V1V3 are discrepant with each other.

Table 7.3. Results of coefficient estimates and hypothesis testing of pairwise comparison with two platforms analyzed at a time.

	$\hat{\alpha}_0$	$\hat{\beta}_0$	$\hat{\beta}_1$	p value of $H_0: (\alpha_0, 1) = (\beta_0, \beta_1)$
V1V2 v.s. V1V3	0.00	-3.09e-5	0.82	3.08e-17
V1V3 v.s. V3V4	0.00	0.00	0.87	1.00
V1V3 v.s. V3V4	7.43e-5	0.00	0.99	0.28

Figure 7.5A – C the relation between estimated mean of abundance of all the bacteria, i.e. $\hat{\xi}_i, i = 1, \cdots, I$, and the corresponding predicted slopes from two platforms, i.e. $A_{i1}, B_{i1}, i = 1, \cdots, I$, and they have the same pattern as it is in Figure 7.4, which helps to confirm that the pairwise comparison works reasonably.

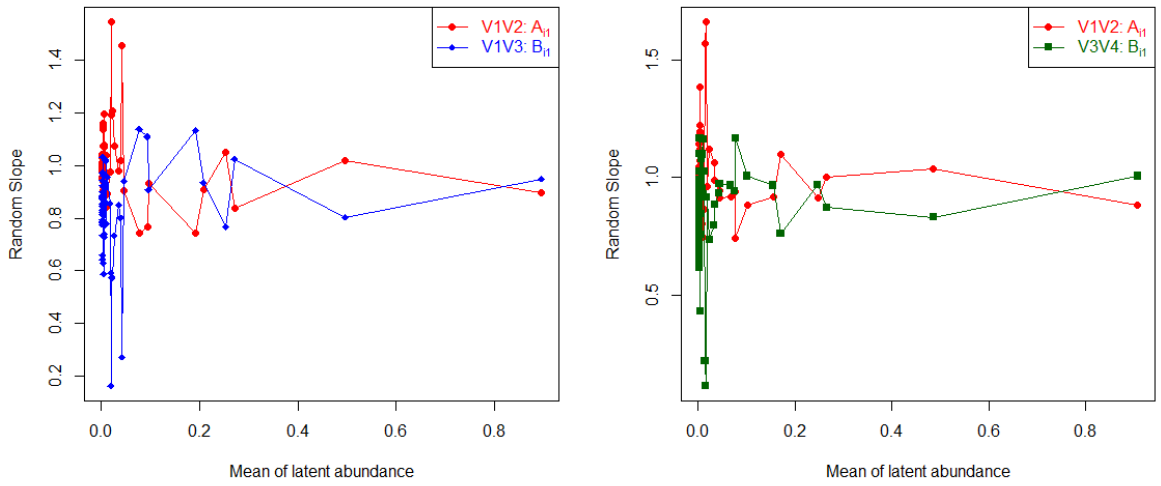

Figure 7.5. A (left) – relation between estimated mean of abundance of all the bacteria, i.e. $\hat{\xi}_i, i = 1, \cdots, I$, and the corresponding predicted slopes from two platforms, i.e. $A_{i1}, B_{i1}, i = 1, \cdots, I$ when comparing V1V2 and V1V3; B (right) – corresponding plot of comparing V1V2 and V3V4.

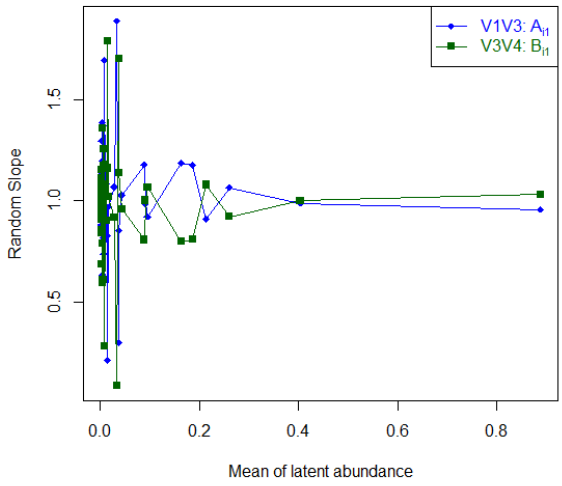

Figure 7.5C. Relation between estimated mean of abundance of all the bacteria, i.e. $\hat{\xi}_i, i = 1, \cdots, I$, and the corresponding predicted slopes from two platforms, i.e. $A_{i1}, B_{i1}, i = 1, \cdots, I$ when comparing V1V3 and V3V4.

Besides, based on the reliability defined in Section 7.9, V1V2, V1V3 and V3V4 have reliabilities 0.40, 0.21 and 0.42, indicating V1V2 and V3V4 have similar overall quality across all the bacteria, while V1V3 gives poor measurements comparatively. Figure 7.6 shows the conditional reliability of each bacterium, i.e. $R^2_{X_i}$, $R^2_{Y_i}$ and $R^2_{Z_i}$ across $\hat{\xi}_i$, which also helps to confirm that V1V3 tends to perform worse than the other two platforms, and measurements would be unreliable when bacteria are rare.

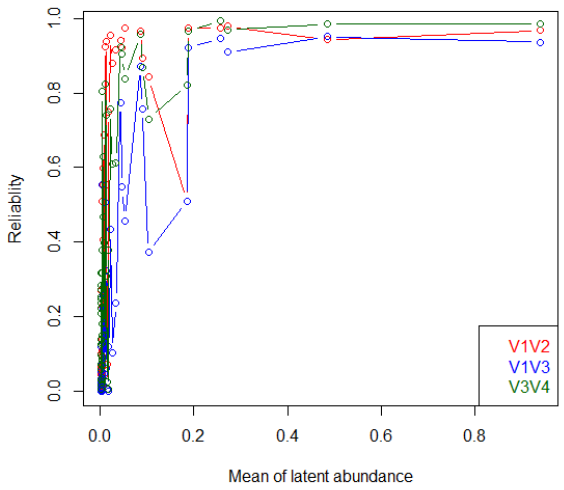

Figure 7.6. Conditional reliabilities of three platforms across each bacteria ordered by the estimated mean abundance $\hat{\xi}_i$.

7.11 Contributions and future work

In this dissertation a model that could compare platforms across large number of genes or bacteria while allowing for heterogeneity for each gene or bacterium was proposed by introducing random effects, while most of the related literatures in terms of platform comparison ignore the individual properties from per gene or bacterium.

Currently our newly developed random effect SEM model is able to handle situations where there are two or three platforms, although the EM algorithm could be easily adjusted to cases where more than three platforms are present, the computing time would increase dramatically, thus it is necessary to work on other algorithms to accelerate the process. More importantly, the performance of MLE should be compared with models with only fixed effects through simulations, which is another reason why a much more time-efficient algorithm is in need.

In parallel with the definition of functional and structural EIV, it is natural that we should consider functional and structural SEM. The model above is clearly structural EIV because $\xi_{ij} \sim N(\xi_i, \sigma_{\xi_i}^2)$, thus the model with $\xi_{ij}'s$ treated as fixed unknown parameters is worth studying. Besides, the statistical inference of method of moments described in Section 7.7 is still undeveloped.

Bibliography

[1] Hemmerle, W. J., & Hartley, H. O. (1973). Computing maximum likelihood estimates for the mixed AOV model using the W transformation.*Technometrics*, *15*(4), 819-831.

[2] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), 1-38.

[3] Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. Biometrika, 58(3), 545-554.

[4] Bollen, K. A. (2014). Structural equations with latent variables. John Wiley & Sons.

[5] Laird, N., Lange, N., & Stram, D. (1987). Maximum likelihood computations with repeated measures: application of the EM algorithm. Journal of the American Statistical Association, 82(397), 97-105.

[6] Davidian, M., & Giltinan, D. M. (1995). Nonlinear models for repeated measurement data (Vol. 62). CRC Press.

[7] Leng, L. (2009). Compound and constrained regression analyses. Stony Brook University. Doctoral Thesis (Advisor: Wei Zhu).

[8] Fuller, W. A. (2009). Measurement error models (Vol. 305). John Wiley & Sons.

[9] Patriota, A. G., Bolfarine, H., & de Castro, M. (2009). A heteroscedastic structural errors-in-variables model with equation error. Statistical Methodology,6(4), 408-423.

[10] Barnett, V. D. (1970). Fitting straight lines-the linear functional relationship with replicated observations. Applied Statistics, 135-144.

[11] Chan, L. K., & Mak, T. K. (1979). Maximum likelihood estimation of a linear structural relationship with replication. Journal of the Royal Statistical Society. Series B (Methodological), 263-268.

[12] Gnatenko, D. V., Zhu, W., & Bahou, W. F. (2008). Multiplexed genetic profiling of human blood platelets using fluorescent microspheres. Thromb Haemost,100(5), 929-936.

[13] Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One, 9(1), e78644.

[14] Xu, X., Zhang, Y., Williams, J., Antoniou, E., McCombie, W. R., Wu, S., ... & Li, E. (2013). Parallel comparison of Illumina RNA-Seq and Affymetrix microarray platforms on transcriptomic profiles generated from 5-aza-deoxy-cytidine treated HT-29 colon cancer cells and simulated datasets. BMC bioinformatics, 14(Suppl 9), S1.

[15] Efron, B., & Efron, B. (1982). The jackknife, the bootstrap and other resampling plans (Vol. 38). Philadelphia: Society for industrial and applied mathematics.

[16] Linnet, K. (1993). Evaluation of regression procedures for methods comparison studies. CLINICAL CHEMISTRY-WASHINGTON-, 39, 424-424.

[17] "Generalized Method of Moments". Retrieved from http://faculty.washington.edu/ezivot/econ583/gmm.pdf.

[18] Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. Econometrica: Journal of the Econometric Society, 1029-1054.

[19] Sun, Z., Kuczek, T., & Zhu, Y. (2014). Statistical calibration of qRT-PCR, microarray and RNA-Seq gene expression data with measurement error models. *The Annals of Applied Statistics*, *8*(2), 1022-1044.

[20] Sun, Z., & Zhu, Y. (2012). Systematic comparison of RNA-Seq normalization methods using measurement error models. *Bioinformatics*, *28*(20), 2584-2591.

[21] Spurgeon, S. L., Jones, R. C., & Ramakrishnan, R. (2008). High throughput gene expression measurement with real time PCR in a microfluidic dynamic array. *PloS one*, *3*(2), e1662.

[22] Chen, J., Agrawal, V., Rattray, M., West, M. A., St Clair, D. A., Michelmore, R. W., ... & Meyers, B. C. (2007). A comparison of microarray and MPSS technology platforms for expression analysis of Arabidopsis. *BMC genomics*,*8*(1), 414.

[23] Arikawa, E., Sun, Y., Wang, J., Zhou, Q., Ning, B., Dial, S. L., ... & Yang, J. (2008). Cross-platform comparison of SYBR® Green real-time PCR with TaqMan PCR, microarrays and other gene expression measurement technologies evaluated in the MicroArray Quality Control (MAQC) study. *BMC genomics*, *9*(1), 328.

[24] Allen, G., Sioutas, C., Koutrakis, P., Reiss, R., Lurmann, F. W., & Roberts, P. T. (1997). Evaluation of the TEOM® method for measurement of ambient particulate mass in urban areas. *Journal of the Air & Waste Management Association*, *47*(6), 682-689.

[25] Wu, X., Berkow, K., Frank, D. N., Li, E., Gulati, A. S., & Zhu, W. (2013). Comparative analysis of microbiome measurement platforms using latent variable structural equation modeling. BMC bioinformatics, 14(1), 79.

[26] Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of molecular biology, 94(3), 441-448.

[27] Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., ... & Volkmer, G. A. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437(7057), 376-380.

[28] Rosey, A. L., Abachin, E., Quesnes, G., Cadilhac, C., Pejin, Z., Glorion, C., ... & Ferroni, A. (2007). Development of a broad-range 16S rDNA real-time PCR for the diagnosis of septic arthritis in children. Journal of microbiological methods, 68(1), 88-93.

[29] Bilonick, R. A., Connell, D. P., Talbott, E. O., Rager, J. R., & Xue, T. (2015). Using structural equation modeling to construct calibration equations relating PM 2.5 mass concentration samplers to the federal reference method sampler.*Atmospheric Environment*, *103*, 365-377.

[30] Janda, J. M., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. Journal of clinical microbiology, 45(9), 2761-2764.

[31] Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K. H., ... & Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature Reviews Microbiology, 12(9), 635-645.

[32] Ahrens, W. H., Cox, D. J., & Budhwar, G. (1990). Use of the arcsine and square root transformations for subjectively determined percentage data. Weed Science, 452-458.

[33] Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. The Annals of Mathematical Statistics, 9(1), 60-62.