Stony Brook University

Statistical Comparison of Measurement Platforms

A Dissertation presented

by

Yuanhao Zhang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

December 2014

Stony Brook University

The Graduate School

Yuanhao Zhang

We, the dissertation committe for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation

Wei Zhu - Dissertation Advisor
Professor, Department of Applied Mathematics and Statistics

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

Xuefeng Wang - 3rd Internal Member
Adjunct Professor, Department of Applied Mathematics and Statistics

Ellen Li - Outside member
Professor, Department of Medicine, Stony Brook University

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Statistical Comparison of Measurement Platforms

by

Yuanhao Zhang

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2014

This thesis proposes a novel statistical method based on the generalized linear errors-in-variables model to compare two measurement platforms with discrete and continuous outcomes respectively. This method overcomes the limitation of the classical platform comparison method with only linear models that can only accommodate two continuous outcome measures. This novel method was applied to model two gene expression measurement platforms: Microarray (continuous) and RNA-Seq (discrete). The comparison result is further validated by differentially expressed gene analysis and biological pathway analysis. The proposed approach would play a significant role: 1) assessing emerging platforms systematically with existing platforms, 2) serving as a foundation to integrate datasets generated from different platforms.

In order to perform platform comparison, a model is built between Microarray and RNA-Seq gene expression profiles based on established distribution assumptions for the purpose of estimating fixed and proportional biases. From both biological and technical view, the variation and dispersion in the measured expression profiles are considered to be gene-specific, which means realistic models of whole genome expression profile datasets contains large number of nuisance parameters and each platform would feature only a limited number of replicates because of the high cost to measure

sample on both platforms. Consequently, substituting those parameters with their common estimates from the limited replicates in the model's likelihood function is often proven unreliable with large variances. Therefore, directly replacing nuisance parameter with estimates from replicates does not lead to appropriate estimates. Additionally, because the number of parameters in model is often tens of thousands, estimating nuisance parameters through their maximum likelihood estimators (MLE) is no longer feasible considering the computational difficulties. In order to overcome above limitations, we further developed a customized estimation method for the proposed generalized linear errors-in-variables model based on unbiased estimating equations (UEE), which yield estimators in analytical form, in lieu of maximum likelihood estimate. Under suitable distribution assumptions of the platforms, the new estimator is proven, theoretically, to converge to the underlying truth with a small bias, which is due to the inherent low count in the discrete platform. The performance of proposed method's was first evaluated by simulated datasets with modest number (three, five and ten) of replicates and subsequently applied to compare published Microarray and RNA-Seq datasets.

Table of Contents

Contents

List of Figures/Tables/Illustrations

List of Figures

List of Tables

List of Abbreviations

AUC ..... Area Under Curve

cDNA ..... complementary DNA

CI ..... Confidence Interval

DEG ..... Differentially Expressed Gene

EIV ..... Errors-in-Variables

FDR ..... False Discovery rate

GLM ..... Generalized Linear Model

GMR ..... Geometric Mean Regression

IPA ..... Ingenuity Pathway Analysis

MLE ..... Maximum Likelihood Estimation

LLN ..... Law of Large Number

LM ..... Linear Model

OTU ..... Operational Taxonomic Unit

OLS ..... Ordinary Least Square

PCA ..... Principle Component Analysis

PCR ..... Polymerase Chain Reaction

ROC ..... Receiver Operating Characteristic

RPKM ..... Reads Per Kilobase of exon per Million reads mapped

SEM ..... Structural Equation Modeling

UEE ..... Unbiased Estimating Equation

Acknowledgements

- Thanks to Prof. Wei Zhu for advising this work.
- Thanks to Prof. Ellen Li for advising in the biological background and providing the dataset motivates this work.
- Thanks to for take the responsibility of dissertation committee.
- Thanks to Prof. Song Wu, Prof. Xuefeng Wang Prof. Pei-fen Kuan for comments during group meetings.
- Thanks to Dr. Xiao Xu, Dr. Jennie Williams, Dr. Eric Antoniou, Dr. W. Richard McCombie, Dr. Song Wu, Dr. Wei Zhu, Dr. Nicholas O. Davidson, Dr. Paula Denoya, and Dr. Ellen Li. In this dissertation, some results are cited from my analysis performed during the process of our published research: "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."
- Also Thanks to Jinmiao Fu and Ruofeng Wen for numerous helpful discussions.

Vita, Publications and/or Fields of Study

Education:

- Stony Brook University, Stony Brook, NY
  Aug. 2010 – Dec. 2014
  - PhD in Applied Mathematics and Statistics (Expected Dec. 2014)
  - MS in Statistics Aug. 2010 Dec. 2012
- Zhejiang University China, Hangzhou, China
  Aug. 2006 – May. 2010
  - BS in Statistics, CKC Honored College in Science

Research & Teaching Experience:

- Stony Brook University, Stony Brook, NY
  PhD Candidate & Research Assistant
  Aug. 2010 – Present
  - **Doctoral thesis research**

    Novel statistical method to compare measurement platforms with errors-in-variables model. Implemented on parallel gene expression profiling datasets generated from Microarray and RNA-Seq technology as groundwork of integrated analysis.
  - **Differentially expressed gene analysis**

    1. Comparison miRNA expression between African and Caucasian American

    2. Gene expression analysis in stem cell, HT-29 colon cancer cell, tumor samples and E.coli strains LF82 and HS.

    3. Performance comparison of existing algorithms resulted in a published paper.
  - **Comparative genome study**
  - **Microbiome analysis with multivariate model**

  - **SNP association study**

- Albert Einstein College of Medicine, Bronx, NY
  Student Trainee
  Jun. 2011 – Aug. 2011

  - **Integrated Study of family and case-control designs in genome-wide association study**
    Development of a new test statistic combining genotype data from different deigns.

- Stony Brook University, Stony Brook, NY

  Teaching Assistant

  Sep. 2011 – May. 2012

  Lectured part of classes in "Probability and Statistics in the Life Science". Also filled in for lecturer of an applied calculus course. Graded homework and exams. Assisted student individually with related problem.

Poster Presentation:

- **Yuanhao Zhang**, Wei Zhu, Leahana Rowehl, Edgar Boedeker, Xuejian Xiong, John Parkinson, Phillip Tarr, Daniel Frank, Grace Gathungu, Ellen Li. "Comparative Transcriptomic Analysis of a Reference Adherent-invasive E. Coli Strain LF82." 2014 Advances in Inflammatory Bowel Diseases, Orlando, FL (2014).

Publications:

Equally Contributed First Author:

- Xiao Xu, **Yuanhao Zhang**, Jennie Williams, Eric Antoniou, W. Richard McCombie, Song Wu, Wei Zhu, Nicholas O. Davidson, Paula Denoya, and Ellen Li. "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 bioinformatics14, no. Suppl 9 (2013): S1.

Co-author:

- Alexander Lee, Navya Kanuri, **Yuanhao Zhang**, Gregory S Sayuk, Ellen Li and Matthew A Ciorba. IDO1 and IDO2 Non-synonymous Gene Variants: Correlation with Crohn's Disease Risk and Clinical Phenotype. PloS one (in press).

- Ellen Li, Ping Ji, Nengtai Ouyang, **Yuanhao Zhang**, Xin Yu Wang, Deborah C. Rubin, Nicholas O. Davidson, Roberto Bergamaschi, Kenneth R. Shroyer, Stephanie Burke, Wei Zhu and Jennie L. Williams. "Differential expression of miRNAs in colon cancer between African and Caucasian Americans: Implications for cancer racial health disparities." International journal of oncology 45, no. 2 (2014): 587-594.

- Matthew Katz, Maryann E. Parrish, Ellen Li, **Yuanhao Zhang**, Wei Zhu, Kenneth Shroyer, Roberto Bergamaschi, and Jennie L. Williams. "The Effect of Race/Ethnicity on the Age of Colon Cancer Diagnosis." Journal of Health Disparities Research and Practice 6, no. 1 (2013): 5.

- Rebecca A. Rowehl, Stephanie Burke, Agnieszka B. Bialkowska, Donald W. Pettet III, Leahana Rowehl, Ellen Li, Eric Antoniou, **Yuanhao Zhang**, Roberto Bergamaschi, Kenneth R. Shroyer, Iwao Ojima and Galina I. Botchkina. "Establishment of Highly Tumorigenic Human Colorectal Cancer Cell Line (CR4) with Properties of Putative Cancer Stem Cells." PloS one 9, no. 6 (2014): e99091.

- Tao Wang, Chang-Yun Lin, **Yuanhao Zhang**, Ruofeng Wen, and Kenny Ye. "Design and statistical analysis of pooled next generation sequencing for rare variants." Journal of Probability and Statistics 2012 (2012): 10.1155/2012/524724.

1 Introduction

It is well known that most, if not all, observed measures are subject to a random error. For example, the error of measuring systolic blood pressure accounts approximately 1/3 of its variations (Carroll et al., 1984). When a variable is measured with a random error that can not be neglected, the variable is often called error-prone variable where its true value is essentially a latent variable (Carroll, 1998).

With the technology improving, human beings keep looking for better instruments, including measurement platforms. Taking the measure of gene expression for example, there was the traditional qPCR, and then the gene Microarray, and now, RNA Sequencing (RNA-Seq) technology. However, the most pressing questions many researchers find themselves asking are, given multiple measurement platforms, are they consistent with each other? Which one is the best choice? Since the another question on how to integrate previous and current data generated by different platforms.

This work is dedicated to better compare gene Microarray and the RNA-Seq platforms two front runners in quantifying whole genome expression profile. In Chapter 3 and 4, a novel approach based on the generalized linear errors-in-variables (EIV) model is proposed to enable the comparison of two platforms with different measurement modalities, in particular, Microarray with its continuous gene expression measures, and RNA-Seq with its count measures. The comparison result is further validated by the differentially expressed gene (DEG) analysis in Chapter 5. Additionally, in Chapter 6, three metagenomics measurement modalities were compared with both simultaneous structural equation modeling (SEM) approach and the pairwise EIV modeling approach. We further compare the SEM and the EIV in terms of platform comparison.

2 Literature Review

2.1 Statistically assessing the agreement between two measurement platforms

Whether caused by the inconsistency from device or its user, no measurement method comes to perfectly represent the true value (Ludbrook, 2010). As a result, the observed measures is always subject to a random measurement error. When new measurement platform emerges, its accuracy is evaluated by its agreement with the existing platform (Martin Bland and Altman, 1986). This section will give a literature review of existing statistical methods that evaluates the agreement between measurement platforms and their application in real datasets.

2.1.1 Measurement error

In practice, as the true value is measured with a random error, the observations of a measurement platform can be modeled as:

$$X_i = \mu_i + error \tag{1}$$

where the X_i is the ith observation of the platform, μ_i is the underlying true value and *error* is the measurement error. In practice, due to the presence of the random variable ϵ_i, an appropriate method to compare two different platforms generally requires repeated measures (Altman and Bland, 1983).

2.1.2 Fitting a straight line

Because the underlying truth will be distributed on the reference line $y = x$ in this case, the most straightforward approach to access the consistency between two platforms is fitting a straight line to observations and compare it to the reference line. Subsequently, the problem becomes to find the optimal fitted line. In this section, we review existing approaches on solving this problem directly from observed pairs of data points.

In (Brace, 1977), the author indicates that it is common in experiments to fit a straight line between two variables X and Y, where both X and Y are subject to measurement errors. When X and Y are essential two measurement methods of the same variable, we are expecting that $y = x$ can be fitted to data. However, there is no basis to determine which one is the independent variable in a tradition ordinary least square regression (OLS) algorithm. Geometrically speaking, after plotting the observations on a $X - Y$ plane, a least square regression with X or Y as the independent variable is to find the straight line that minimizes the sum of vertical or horizontal distances from observed data points to it. It implies, taking X as independent variable as an example, the best fitted value for latent truth is the observation itself. The same goes for using Y as independent variable. Both types of OLS regressions are known to be inconsistent because of the measurement errors. In order to reduce the biases between OLS regression line and the real line, the author proposes two methods: 1) to minimize the perpendicular distance and 2) to minimize the weighted sum square of the vertical and horizontal distances (see Figure 1), where the corresponding weight is the $\frac{1}{SE_y}$ and $\frac{1}{SE_x}$.

In (Ludbrook, 1997), the problem of fitting a straight line to two different measures is revisited and further elaborated. In this paper, the comparison of two measurement platforms is summarized as identification of two types of systematic biases: the fixed and the proportional bias. When there is neither type of biases, the underlying true values distributed on the line $y = x$. Whereas when the underlying true values distributed on the line $y = a + bx$ ($a \neq 0$ and $b \neq 1$), a is defined as the fixed bias and b is defined as the proportional bias. In this sense, the correlation coefficient is no longer an appropriate quantification of platforms' consistency given it provides no information whether the data points scatter around $y = x$ or any other straight line. In addition, estimates of both types of biases cannot be computed through OLS algorithm because of the in consistent stated in the paragraph above. Therefore, the author proposes to apply an ordinary least product (OLP) regression instead of OLS regression. OLP regression aims to minimize the product of the vertical and horizontal distance, which be viewed as another interpretation of the weighted sum square approach described in (Brace, 1977). A bootstrap approach of estimating the confidence intervals for both types of biases is also illustrated. Therefore a hypothesis test of whether the platforms are consistent can be translated to whether the

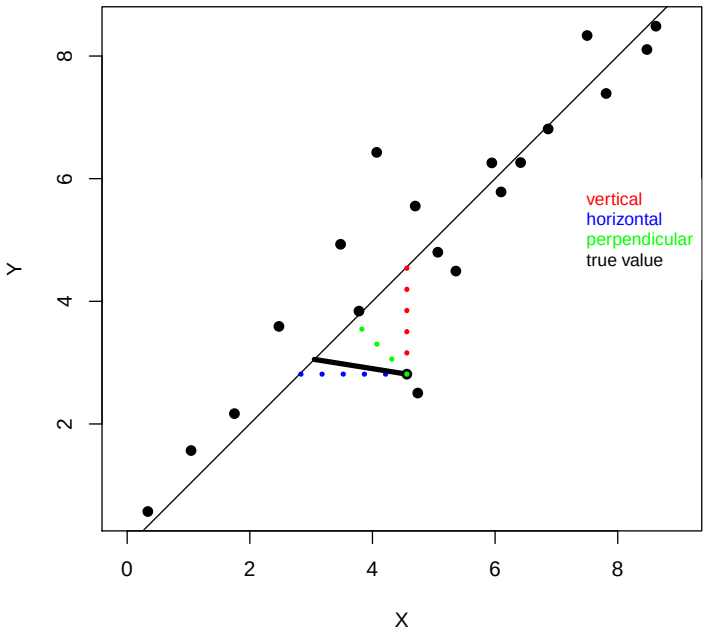

Figure 1: Different types of distance

A demonstration of different distances from observed data point to a straight line.

confidence interval includes $a = 0$ and $b = 1$ at a given confident level.

Besides the regression based methods, it is also proposed to use the median of all intercepts generated by connecting any two data points as the slope for the fitted line (Passing and Bablok, 1983). Another method is grouping data points into two clusters, and use the line connecting the cluster centroids (Wald, 1940). However, those methods often suffer from large estimating bias when the underlying truth follows a unimodal distribution.

Different from methods mentioned above, when we assume the measurement errors follows a distribution, fitting the straight line becomes a problem of estimating a and b through building a model between platforms with EIV (two platforms) and SEM(three or more platforms), which is reviewed in detail later in this chapter.

2.1.3 Consistency of measuring changes

When there is not enough information to model both platforms, obtaining a consistent estimate of the underlying true line is challenging especially when there is no repeated measures. However, the level of platforms' consistency can be also indirectly assessed by the outcome of the measurement changes under different conditions. For example, when compare gene expression profiles measured by Microarray and RNA-Seq, the outcome could be expressed as the overlap rate of DEG (Zhao et al., 2014). This approach has been applied to a number of comparisons between parallel RNA-Seq and Microarray datasets. A pioneer work of human kidney and live DEG comparison is reported in (Marioni et al., 2008), where same RNA samples were measured with both platforms and the DEGs are identified by empirical Bayes T statistics (Smyth et al., 2004) and a Poisson model respectively. The result shows 81% of DEGs in Microarray is overlapped with the DEGs in RNA-Seq. Similar work has also been done in other datasets (Mooney et al., 2013; Xu et al., 2013; Zhao et al., 2014) with comparable results of overlap rates.

2.1.4 Comparing platform by reliability

Measurement reliability is the quantification of the degree of reproducibility over replicates. Although it does not reflect the existence of the two types of systematic biases, a valid measurement platform naturally require a certain level of reliability to be reliable in experiments. Assessing the reliabilities of platforms is also a significant consideration in platform comparison. Specifically, when applying to measurement error models, the reliability is defined as the square of correlation between measurements and the latent truth (Allen and Yen, 2001).

2.1.5 Statistical calibration

In platform comparison, sometimes a given platform may be considered as the 'gold standard', but its application is limited by other factors. In contrast, the alternative methods have a relative low accuracy but better feasibility. For example, measuring gene expression with reverse transcription polymerase chain reaction is considered as the most accurate method, but testing genome-wide expression profile with this method may not be feasible because of the limitation of RNA samples. In contrast, Microarray and RNA-Seq yield information on thousands of genes in every single run. A

comparative statistical calibration is an important application of platform comparison, where measurements are adjusted according to the most accurate platform. (Osborne, 1991).

2.2 Linear EIV model in comparing platforms

2.2.1 Introduction

EIV model, or measurement error model, aims to model the relationship between latent variables. In EIV, the observations are essentially surrogates with random errors (Liang, 2000). The classical linear EIV model with normality assumption can be written as

$$\begin{aligned} x_i &= \xi_i + \delta_i \\ y_i &= \eta_i + \epsilon_i \\ \eta_i &= a + b\xi_i \\ \delta_i &\overset{i.i.d}{\sim} N(0, \sigma_\delta^2) \\ \epsilon_i &\overset{i.i.d}{\sim} N(0, \sigma_\epsilon^2) \end{aligned} \tag{2}$$

where ξ_i and η_i are the latent truth for observations x_i and y_i. δ_i and ϵ_i are the corresponding measurement errors following normal distributions.

If ξ_i are fixed constants, the model (2) is called functional model, if $\xi_i \sim N(\mu_\xi, \tau^2)$, it is called a structural model. The measurement errors are considered to be a series of independent random variables both within and across the platforms. In additional, under the structural model, the variables of latent truths and measurement errors together are also considered to be independent variables. It is well known that under this model, the OLS estimate of the slope will be inconsistent and converge to $\frac{b}{1+\sigma_\delta^2/\sigma_\xi^2}$ (Greene, 2003), also known as the attenuation effect.

Let us denote the true value for platform A to be ξ and for platform B to be η. After transforming into the same dimension, a perfect consistency between platforms requires $\eta = a + b\xi$, and $a = 0$, $b = 1$. If $a \neq 0$, a is called fixed bias; if $b \neq 1$, b is called proportional bias (Ludbrook, 2010). Therefore, under normality assumptions, model (2) formulates the problem of comparing platforms via fitting a straight line and leads to the estimate of the fixed and proportional biases as discussed in Section 2.1.2. When both ξ_i and η_i are observable, a linear regression model will give the estimates

of both types of biases. However, instead of ξ and η, what we observe are surrogates X and Y with measurement errors and the OLS estimates of a and b will be biased. The true in the EIV model, or the EIV line, is bounded by the slope of OLS regression $y \sim x$ and inverse slope of OLS regression $x \sim y$, which is known as the Fricsh bound (Chen et al., 2007).

2.2.2 Estimate of slope

In this section, we review different methods of estimating slope of the EIV line because once the slope is estimated, the intercept can be computed with sample means from each platform. Those methods can be classified into three categories: 1) likelihood approach, 2) method of moments and 3) robust estimation.

Likelihood approach. In functional model from (2), the likelihood is

$$L(x_i, y_i; \xi_i, a, b, \sigma_\delta, \sigma_\epsilon) = \Pi_i \frac{1}{2\pi\sigma_\delta\sigma_\epsilon} exp\{-\frac{(x_i - \xi_i)^2}{2\sigma_\delta^2} - \frac{(y_i - a - b\xi_i)^2}{2\sigma_\epsilon^2}\}$$

and in structural model from (2), the likelihood function can be derived from the joint distribution:

$$(x_i, y_i) \overset{i.i.d}{\sim} N\left(\begin{pmatrix} \mu_\xi \\ a + b\mu_\xi \end{pmatrix} \begin{pmatrix} \tau^2 + \sigma_0^2 & b\tau^2 \\ b\tau^2 & b^2\tau^2 + \sigma_0^2 \end{pmatrix}\right).$$

Denote the density function for above bivariate normal distribution as

$$p(x_i, y_i; \mu_\xi, \tau, a, b, \sigma_\delta, \sigma_\epsilon)$$

the likelihood function is

$$L(x_i, y_i; \mu_\xi, \tau, a, b, \sigma_\delta, \sigma_\epsilon) = \Pi_i p(x_i, y_i; \mu_\xi, \tau, a, b, \sigma_\delta, \sigma_\epsilon)$$

If $\sigma_\epsilon^2/\sigma_\delta^2$ is known, there is an explicit solution using likelihood method is

$$b^2 S_{xy} + b(\frac{\sigma_\epsilon^2}{\sigma_\delta^2} S_{xx} - S_{yy}) - \frac{\sigma_\epsilon^2}{\sigma_\delta^2} S_{xy}$$

where S_{xx}, S_{yy} and S_{xy} are sample variance and covariance. The obtained line is equivalent to a weighted total least square regression (Xu et al., 2014):

$$min_{a,b}\{\frac{SSE_X}{\sigma_\delta^2} + \frac{SSE_Y}{\sigma_\epsilon^2}\}.$$

where SSE is the sum square of error.

However, the variance ratio $k = \sigma_\epsilon^2/\sigma_\delta^2$ is generally unknown in a platform comparison problem, hence the MLE equation system is underdetermined (Lindley, 1947). In this case, additional constraints or information are added to model (2) to obtain a MLE.

In (Barnett, 1970), the author discusses on fitting a functional linear EIV model with replications. If there are J replications available for each sample i, and the ratio of measurement error variances remains the same across samples as k. The model can be written as

$$\begin{aligned} x_{ij} &= \xi_i + \delta_{ij} \\ y_{ij} &= \eta_i + \epsilon_{ij} \\ \eta_i &= a + b\xi_i \\ \delta_{ij} &\overset{i.i.d}{\sim} N(0, \sigma_i^2) \\ \epsilon_{ij} &\overset{i.i.d}{\sim} N(0, k\sigma_i^2), \end{aligned}$$

where $i = 1, 2, ..., I$ denote different subjects and $j = 1, 2, ..., J$ denote replications. Let $x_{i.} = \Sigma_j x_{ij}$, $\bar{x}_{i.} = \Sigma_j \frac{x_{ij}}{J}$ and $y_{i.} = \Sigma_j y_{ij}$, $\bar{y}_{i.} = \Sigma_j \frac{y_{ij}}{J}$. Under this model, because of replicates, MLE can be obtained from solving

$$\begin{aligned} &\Sigma_i \frac{(\bar{y}_{i.} - a - b\hat{\xi}_i)}{k\hat{\sigma}_i{}^2} = 0 \\ &\Sigma_i \frac{\hat{\xi}_i(\bar{y}_{i.} - a - b\hat{\xi}_i)}{k\hat{\sigma}_i{}^2} = 0 \\ &\hat{\sigma}_i{}^2 = \frac{\Sigma_j(x_{ij} - \hat{\xi}_i)^2 + \frac{1}{k}\Sigma_j(y_{ij} - a - b\hat{\xi}_i)^2}{2J} \\ &(\bar{x}_{i.} - \hat{\xi}_i) + b(\bar{y}_{i.} - a - b\hat{\xi}_i)/k = 0 \\ &k = \Sigma_i\Sigma_j \frac{(y_{ij} - a - b\hat{\xi}_i)^2 \hat{\sigma}_i{}^{-2}}{IJ}. \end{aligned}$$

Because there is no explicit solution, numeric methods need to be implemented to solve the equation system. Although it is proposed in this paper to add an additional constraint that all subject has equal measurement error variances (σ_i=τ) to reduce the number of score functions, it is still required to compute every single $\hat{\xi}_i$. When the subject number I is large and the

replication number J is small, numerical solution of proposed method is not feasible.

In (Chan and Mak, 1979), under the constraint that the variance of measurement error remains the same across samples, a structural linear EIV model with replications is proposed to reduce the number of parameters in the functional model above,

$$\begin{aligned}
x_{ij} &= \xi_i + \delta_{ij} \\
y_{ij} &= \eta_i + \epsilon_{ij} \\
\eta_i &= a + b\xi_i \\
\delta_{ij} &\overset{i.i.d}{\sim} N(0, \sigma_\delta^2) \\
\epsilon_{ij} &\overset{i.i.d}{\sim} N(0, \sigma_\epsilon^2) \\
\xi_i &\overset{i.i.d}{\sim} N(\mu, \tau^2).
\end{aligned}$$

With this model, let

$$\begin{aligned}
s_{xx} &= \Sigma_i \frac{(\bar{x}_{i.} - \bar{x}_{..})^2}{I} \\
s_{yy} &= \Sigma_i \frac{(\bar{y}_{i.} - \bar{y}_{..})^2}{I} \\
s_{xy} &= \Sigma_i \frac{(\bar{x}_{i.} - \bar{x}_{..})(\bar{y}_{i.} - \bar{y}_{..})}{I} \\
w_{xx} &= \Sigma_i \Sigma_j \frac{(x_{ij} - \bar{x}_{i.})^2}{IJ} \\
w_{yy} &= \Sigma_i \Sigma_j \frac{(y_{ij} - \bar{y}_{i.})^2}{IJ} \\
t_{xx} &= \Sigma_i \frac{x_{ij} - \bar{x}_{..})^2}{I} \\
t_{yy} &= \Sigma_i \frac{(y_{ij} - \bar{y}_{..})^2}{I}.
\end{aligned}$$

The estimate of slope can be obtained from

$$
\begin{aligned}
&m_0 b^4 + m_1 b^3 + m_2 b^2 + m_3 b + m_4 = 0 \\
&m_0 = (J-1) s_{xx} s_{xy} t_{xx} \\
&m_1 = J s_{xx}^2 w_{yy} - (J-1) s_{xy}^2 t_{xx} - (J-1) s_{xx} s_{yy} t_{xx} - J s_{xy}^2 w_{xx} \\
&m_2 = (3J-1)(s_{xy} s_{yy} w_{xx} - s_{xy} s_{xx} w_{yy}) \\
&m_3 = J s_{xy}^2 w_{yy} + (J-1) s_{xy}^2 t_{yy} + (J-1) s_{xx} s_{yy} t_{yy} - J s_{yy}^2 w_{xx} \\
&m_4 = -(J-1) s_{xy} s_{yy} t_{yy}.
\end{aligned}
$$

Method of moments. In (Kendall et al., 1946), it is shown that structural errors-in-variables model can be formulated to an underdetermined equation system with first and second order moments. Different to likelihood approach, additional information can be gained from including third or higher order of moments. A review of method of moments is shown in (Gillard, 2006) which lists an estimator based the property of odd order moments of symmetric distribution: given the (2m+1) order central moment exists,

$$\frac{1}{I} \Sigma_i [y_i - \bar{y} - b(x_i - \bar{x})]^{(2m+1)} = 0$$

is a consistent estimator (Scott, 1950). In additional, estimators based on 4th order moment is discussed in (Cragg, 1997). The common problem of estimator with high order moments is the relative low efficiency comparing the likelihood based method.

Robust estimate. In (Zamar, 1989), the author gives a robust and non-parametric approach to obtain the equation of the underlying line. It is argued the OLS estimate can be rewritten as a minimization problem:

$$\min_{a,b} \{\Sigma_i (y_i - a - b x_i)^2\}$$

which is equivalent to orthonormal M estimates

$$
\begin{aligned}
&\min_{a,b} \{\Sigma_i f(\frac{y_i - a - b x_i}{S_i})\} \\
&f(x) = x^2 \\
&S_i = 1.
\end{aligned}
$$

Therefore a robust estimate can be written as

$$\min_{a,b}\{\Sigma_i f_1(\frac{y_i - a - bx_i}{S_i})\}$$

where f_1 is a robust featured loss function like Tukey's loss function. S_i is M scale defined by

$$\{\Sigma_i f_2(\frac{y_i - a - bx_i}{S_i})\} = \Delta$$

where f_2 can be any reasonable loss function.

Although such estimator does not guarantee consistency, it still fits a line that is close to the pattern in data and potentially has a better performance than OLS regression in some cases.

Special cases. Due to the general difficult in specifying the variance ratio $k = \sigma_\epsilon^2/\sigma_\delta^2$, some authors propose to use special variance ratio when no other information available; e.g. in (Wong, 1989), the author assumes $\sigma_0 = \sigma_\delta = \sigma_\epsilon$ and consequently the estimated line minimizes the sum square of perpendicular distances from data points, which is also known as orthogonal regression (OR). Another case that has already been mentioned in the above sets $\sigma_\epsilon^2/\sigma_\delta^2 = \frac{S_{yy}}{S_{xx}}$ and the resulting method is known as the geometric mean regression (GMR).

2.2.3 Application in platform comparison

Because of the challenges mentioned in above sections, it is often difficult to build a model that fits the real platform structure and satisfies the constraints in estimating methods. In practice, most of comparisons have been performed with the OR and GMR regressions due to its advantages discussed in (Brace, 1977) and (Ludbrook, 1997). In (Ludbrook, 2010), the author lists a series of studies that compare various platforms with those special EIV models. Examples are:

1. comparing the methods of concentration measurements like continuous flow versus manual flame photometry for sodium as well as atomic absorption versus continuous-flow for calcium (Cornbleet and Gochman, 1979).

2. comparing analytical methods used in Brazilian prospects for copper,

gold and iron (Castilho, 2004).

3. comparing platelet genetic biomarker quantification fluorescent with Microspheres and PCR platforms (Huang et al., 2013).

2.3 SEM in comparing platforms

In this section, we review using SEM to compare platforms in a context of comparing three platform, but comparison of more platforms follows the same principle.

2.3.1 Introduction

SEM is a very general statistical modeling approach that combines regression and network together (Hox and Bechger, 1998). In comparing 3 or more platforms, the model can be written as a confirmatory factor analysis where the latent factor is the measure without any error with the data on each platform centered before analysis (Wu et al., 2013). Following conventional parameterization setting (Bollen, 1998), let ξ denote the latent true value, the SEM model to compare three centered datasets can be written as

$$\begin{aligned} x_i &= \xi_i + \delta_i \\ y_i &= b_y \xi_i + \epsilon_i \\ z_i &= b_z \xi_i + \omega_i \end{aligned}$$

or in matrix form:

$$\mathbf{T} = \mathbf{b}'\xi + \mathbf{r}$$

where $\mathbf{T} = \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix}$, $\mathbf{b}' = \begin{pmatrix} 1 \\ b_y \\ b_z \end{pmatrix}$ and $\mathbf{r} = \begin{pmatrix} \delta_i \\ \epsilon_i \\ \omega_i \end{pmatrix}$. The path diagram is shown is Figure 2. We further assume that all errors are normally distributed and independent, the latent variable $\xi_i \sim N(\mu, \tau^2)$ and is independent from measurement errors.

Covariance matrix $\mathbf{\Lambda} = E(\mathbf{T}\mathbf{T}') = \mathbf{b}\mathbf{b}'\tau^2 + \mathbf{R}$ where $\mathbf{R} = diag(\sigma_\delta^2, \sigma_\epsilon^2, \sigma_\omega^2)$. Using sample covariance matrix $\mathbf{S}$ to equate $\mathbf{\Lambda}$, the estimator can be obtained

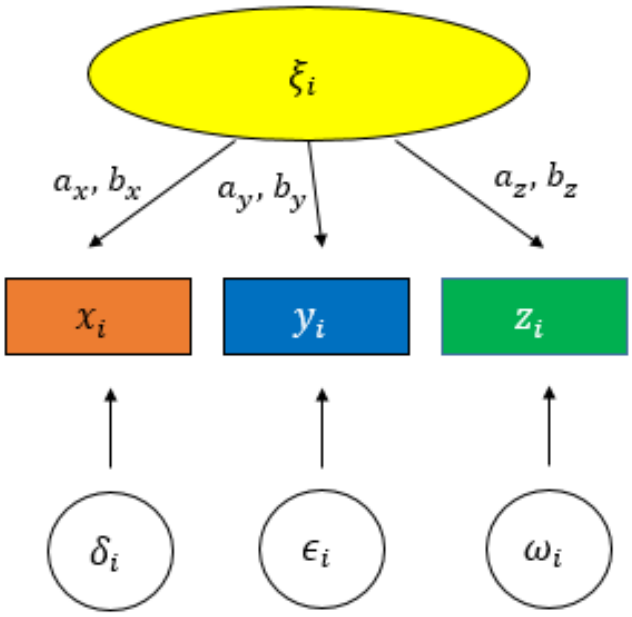

Figure 2: Demonstration of SEM path diagram for measurement error model

as (Wu et al., 2013)

$$\hat{b}_y = \frac{S_{23}}{S_{13}}\ \hat{b}_z = \frac{S_{23}}{S_{12}}\ \hat{\tau}^2 = \frac{S_{12}S_{13}}{S_{23}}$$
$$\hat{\sigma}_\delta^2 = S_{11} - \hat{\tau}^2\ \hat{\sigma}_\epsilon^2 = S_{22} - \hat{\tau}^2\ \hat{\sigma}_\omega^2 = S_{33} - \hat{\tau}^2.$$

2.3.2 Platform reliability and calibration

Besides estimating the systematic biases, SEM can also be applied to computing platform reliabilities and statistical calibration with three or more platforms.

Reliability. Comparison of reliability may be helpful when the platforms show little bias with each other. It can be easily shown that the squared correlation between measurements and the latent truth can be estimated by

$$\rho_{x,\xi}^2 = \frac{\rho_{x,y}\rho_{x,z}}{\rho_{y,z}}\ \rho_{y,\xi}^2 = \frac{\rho_{x,y}\rho_{y,z}}{\rho_{x,z}}\ \rho_{z,\xi}^2 = \frac{\rho_{y,z}\rho_{x,z}}{\rho_{x,y}}\ .$$

$(\rho_{x,\xi}^2, \rho_{y,\xi}^2, \rho_{z,\xi}^2)$ quantifies reliabilities for three platforms in measurement theory (Allen and Yen, 2001) under platform comparison scenario.

Calibration. As discussed in Section 2.1, sometimes one platform is considered to generate more accurate measures of the latent variable and calibration on other platforms are desired. The estimated parameters from the SEM model on measurement platforms can be considered as proportional biases to calibrate observed measurements and the fixed biases can be estimated from sample means. Without losing generosity, we let platform x denote the 'reference' measures and y, z as alternative platforms. Under regularity conditions, the calibration method is shown in (Zhao et al., 2014),

$$x_i = \xi_i + \delta_i$$
$$y_i = a_y + b_y\xi_i + \epsilon_i$$
$$z_i = a_z + b_z\xi_i + \omega_i$$

The parameters except a_y and a_z can be obtained with a SEM without intercept after centering the datasets, while the estimates of intercepts can

be obtained from the platform means and estimated $\hat{b}_y$ and $\hat{b}_z$. With the estimated parameters, the calibration approach can be written as

$$\mu_i = \frac{\frac{x_i}{\hat{\sigma}_\delta^2} + \frac{\hat{b}_y(y_i - \hat{a}_y)}{\hat{\sigma}_\epsilon^2} + \frac{\hat{b}_z(z_i - \hat{a}_z)}{\hat{\sigma}_\omega^2}}{\frac{1}{\hat{\sigma}_\delta^2} + \frac{\hat{b}_y^2}{\hat{\sigma}_\epsilon^2} + \frac{\hat{b}_z^2}{\hat{\sigma}_\omega^2}}, i \in \mathcal{A}$$

$$\mu_i = \frac{\frac{\hat{b}_y(y_i - \hat{a}_y)}{\hat{\sigma}_\epsilon^2} + \frac{\hat{b}_z(z_i - \hat{a}_z)}{\hat{\sigma}_\omega^2}}{\frac{\hat{b}_y^2}{\hat{\sigma}_\epsilon^2} + \frac{\hat{b}_z^2}{\hat{\sigma}_\omega^2}}, i \in \mathcal{B} - \mathcal{A}$$

where $\mathcal{A} = \{i | subject\ i\ measured\ by\ platform\ X, Y, Z\}$ and
$\mathcal{B} = \{i | subject\ i\ measured\ by\ platform\ Y, Z\}$.

2.3.3 Application to platform comparison

In (Wu et al., 2013), the authors apply SEM to compute reliabilities of four modalities of quantifying the abundance of Clostridium GroupXIVa: Sanger, 454 pyrosequencing on V1-V3, V3-V5 region of 16s ribosomal RNA and qPCR. The reliabilities are calculated and the 454 pyrosequencing windows feature the highest correlations with latent variable.

Application of SEM on calibration is to adjust Microarray and RNA-Seq with qRT-PCR as reference because it is regarded as "most reliable" in gene expression measurements. In (Zhao et al., 2014), the author implements this calibration method on qRT-PCR, RNA-Seq and Microarray gene expression data of the human brain reference RNA sample and the human universal reference RNA sample. However, only 477 genes are measured with all three platforms while both Microarray and RNA-Seq measures thousands of genes, therefore it is necessary to assume homogeneity of variance in measurement errors across all genes to apply the calibration to genes that are not measured by qRT-PCR.

3 Extend generalized linear EIV model to platform comparison

In Chapter 2, we reviewed existing methods to compare two measurement platforms with a linear EIV model where both measures are required to be continuous. However, in some platform, the observed measurement sometimes may be discrete variables. For example, when sequencing-based technology RNA-Seq, which yields discrete counts, is compared to Microarray, which yields continuous intensities, linear EIV models are no longer appropriate. In this section, we extend the linear EIV model to generalized linear EIV (GEIV) model to compare two platforms with a discrete (count) with aand continuous (color intensity) outputs.

3.1 Introduction to generalized linear EIV model

GEIV can be viewed as the counterpart of generalized linear model in the EIV family where the model is fitted with considering a random error in the independent variables. In canonical form (McCullagh and Nelder, 1989), the GEIV in platform comparison can be written as

$$\begin{aligned} p_Y(y_i; a, b, \lambda_i) &= exp\{\frac{y_i f(\lambda_i) - \psi_1(\lambda_i)}{\psi_2(\phi)} + c(y_i, \phi)\} \\ x_i &= \mu_i + \delta_i \\ f(\lambda_i) &= a + b\mu_i \end{aligned} \tag{3}$$

where p_Y is the density function for y_{ij} with respect to a probability measure; λ_i and μ_i is the latent variable for platform Y and X; $f(\cdot)$ is the link function and $i = 1, 2, ..., I$ is the index for observed measures. It is also assumed the measurement error in X follows a normal distribution $\delta_i \sim N(0, \sigma_\delta^2)$ and is independent from μ_i.

In platform comparison, $f(\cdot)$ can be considered as the function transfer the measurement Y to the same unit with measurement X and it follows a and b remains to be fixed and proportional biases in that unit. If each μ_i is

a constant, the model is a functional EIV model, otherwise it is a structural EIV model.

3.2 Estimation of intercept and slope with known $\Omega = \frac{\psi_2(\phi)}{\sigma_\delta^2}$

Functional model. Under model (3), the likelihood can be expressed as

$$L = \Pi_i \frac{1}{\sqrt{2\pi}\sigma_\delta} exp\{\frac{y_i f(\lambda_i) - \psi_1(\lambda_i)}{\psi_2(\phi)} + c(y_i, \phi) - \frac{(x_i - \mu_i)^2}{2\sigma_\delta^2}\}$$

and

$$l = log(L) = -\frac{I}{2} log(2\pi\sigma_\delta^2) + \Sigma_i \{\frac{y_i f(\lambda_i) - \psi_1(\lambda_i)}{\psi_2(\phi)} + c(y_i, \phi) - \frac{(x_i - \mu_i)^2}{2\sigma_\delta^2}\}.$$

It is discussed in (Stefanski, 1988) that a sufficient condition that the score functions build a well-defined equation system is $\Omega = \frac{\psi_2(\phi)}{\sigma_\delta^2}$ is known, which reduces to the variance ratio under a linear model.

Under a constant model, the parameters of interest is a and b, whereas $I + 2$ nuisance parameters are also present in model. The MLE approach suffers from the computational challenge in non-linear maximization problem especially when I is large.

An alternative approach to MLE is proposed in (Stefanski, 1988). Let $\Delta_i = x_i + y_i \Omega b$ and it is trivial to show Δ_i is a sufficient statistic of μ_i. Therefore, $p_Y(y_i|\Delta_i; a, b, \sigma_\delta^2, \mu_i, \phi) = p_Y(y_i|\Delta_i; a, b, \sigma_\delta^2, \phi)$ and $H(x_i, y_i, \boldsymbol{\theta}) = p_Y(y_i|\Delta)$ is a implicit function of θ, where $\boldsymbol{\theta} = (a, b, \phi, \sigma_\delta)^T$. It is easy to show $E_Y(\frac{\partial H}{\partial \boldsymbol{\theta}}) = 0$. With the notations, a conditional score is defined (Lindsay, 1982) as the solution of $\Sigma_i h(x_i, y_i, \boldsymbol{\theta}) = 0$, where $h = \frac{\partial H}{\partial \boldsymbol{\theta}}$ with a general form:

$$\begin{aligned}
\Sigma_i \frac{y_i + E(y_i|\Delta_i)}{\psi_2(\phi)} &= 0 \\
\Sigma_i \frac{[y_i + E(y_i|\Delta_i)]t(\Delta_i)}{\psi_2(\phi)} &= 0 \\
\Sigma_i r(y_i, x_i, \boldsymbol{\theta}) - \Sigma_i E(r(y_i, x_i, \boldsymbol{\theta})|\Delta_i) &= 0 \\
\frac{\psi_2(\phi)}{\sigma_\delta^2} &= \Omega
\end{aligned}$$

where the form $r(\cdot)$ is determined by distribution of y_i and link function. $t(\Delta_i)$ is a function not depending on x_i and y_i given Δ_i.

Structural model. Under model (3), let $\boldsymbol{\theta} = (a, b, \phi, \sigma_\delta)^T$, $\Omega = \frac{\psi_2(\phi)}{\sigma_\delta^2}$ and $\Delta_i = x_i + y_i \Omega b$. Denote density function of μ_i as $p(\mu_i)$. The likelihood is

$$\begin{aligned} & L(\boldsymbol{\theta}, \mu_i; x_i, y_i) \\ & = \Pi_i \int \frac{1}{\sqrt{2\pi}\sigma_\delta} exp\{\frac{y_i f(\lambda_i) - \psi_1(\lambda_i)}{\psi_2(\phi)} + c(y_i, \phi) - \frac{(x_i - \mu_i)^2}{2\sigma_\delta^2}\} p(\mu_i) d\mu_i \\ & = \Pi_i \int p(x_i, y_i; \boldsymbol{\theta}, \mu_i) p(\mu_i) d\mu_i \,. \end{aligned}$$

Denote $l(x_i, y_i; \boldsymbol{\theta}, \mu_i) = log(L(x_i, y_i; \boldsymbol{\theta}, \mu_i))$, it follows that

$$\frac{\partial}{\partial \boldsymbol{\theta}} l(x_i, y_i; \boldsymbol{\theta}, \mu_i) = E[\frac{\partial}{\partial \boldsymbol{\theta}} p(x_i, y_i; \boldsymbol{\theta}, \mu_i) | x_i, y_i] = E[\frac{\partial}{\partial \boldsymbol{\theta}} p(x_i, y_i; \boldsymbol{\theta}), \mu_i | \Delta_i] \,.$$

A conditional score is defined similarly to the functional model (Stefanski, 1988) with a general form

$$\begin{aligned} \Sigma_i \frac{y_i + E(y_i|\Delta_i)}{\psi_2(\phi)} &= 0 \\ \Sigma_i \frac{[y_i + E(y_i|\Delta_i)] E(\mu_i|\Delta_i)}{\psi_2(\phi)} &= 0 \\ \Sigma_i r(y_i, x_i, \boldsymbol{\theta}) - \Sigma_i E(r(y_i, x_i, \boldsymbol{\theta})|\Delta_i) &= 0 \\ \frac{\psi_2(\phi)}{\sigma_\delta^2} &= \Omega \,. \end{aligned}$$

3.3 Estimation of intercept and slope in platform comparison

In a platform comparison scenario, Ω is unknown and the estimating equation system in above section is underdetermined. However, with replications $j = 1, 2, ..., J$, we can derive an alternative method to estimate.

$$\begin{aligned} p_Y(y_{ij}; a, b, \lambda_i) &= exp\{\frac{y_{ij} f(\lambda_i) - \psi_1(\lambda_i)}{\psi_2(\phi)} + c(y_{ij}, \phi)\} \\ x_{ij} &\sim N(\mu_i, \sigma_\delta^2) \\ f(\lambda_i) &= a + b\mu_i \Leftrightarrow \mu_i = \frac{f(\lambda_i) - a}{b} \,. \end{aligned} \tag{4}$$

Functional model. The log likelihood function is

$$l = -\frac{IJ}{2} log(2\pi\sigma_\delta^2) + \Sigma_i\Sigma_j\{\frac{y_{ij}f(\lambda_i) - \psi_1(\lambda_i)}{\psi_2(\phi)} + c(y_{ij},\phi) - \frac{(x_{ij}-\mu_i)^2}{2\sigma_\delta^2}\}\,.$$

It can be derived the score functions are

$$\begin{aligned}
\frac{\partial}{\partial a}l &= -\Sigma_i\Sigma_j\frac{x_{ij}-\mu_i}{\sigma_\delta^2} = 0 \\
\frac{\partial}{\partial b}l &= -\Sigma_i\Sigma_j\frac{(x_{ij}-\mu_i)\mu_i}{b\sigma_\delta^2} = 0 \\
\frac{\partial}{\partial \lambda_i}l &= \Sigma_j\frac{y_{ij}f'(\lambda_i) - \psi_1'(\lambda_i)}{\psi_2(\phi)} + \Sigma_j\frac{(x_{ij}-\mu_i)f'(\lambda_i)}{b\sigma_\delta^2} = 0 \\
\frac{\partial}{\partial \sigma_\delta}l &= -IJ\frac{1}{\sigma_\delta} + \Sigma_i\Sigma_j\frac{(x_{ij}-\mu_i)^2}{b\sigma_\delta^3} = 0 \\
\frac{\partial}{\partial \phi}l &= -\Sigma_i\Sigma_j\psi_2'(\phi)\frac{y_{ij}f(\lambda_i) - \psi_1(\lambda_i)}{\psi_2^2(\phi)} + \Sigma_i\Sigma_j\frac{\partial}{\partial\phi}c(y_{ij},\phi) = 0 \\
\mu_i &= \frac{f(\lambda_i)-a}{b}\,.
\end{aligned} \tag{5}$$

Equation system (5) also suffers from computational challenges in nonlinear maximization. However, in platform comparison, our primary objective is to estimate a and b. Therefore unbiased estimating equations (UEE) (Cox, 1993), which are functions of statistics and parameters with expectation equal to 0, can be constructed to bypass the nuisance parameters. Denote $\mathbf{y_i} = \{y_{ij}|j = 1, 2, ..., J\}$ and $S_{xx} = \Sigma_i\frac{(x_{ij}-\bar{x}_{i.})^2}{J-1}$. Assume for some functions $T_1(\cdot\,;\boldsymbol{\theta})$ and $T_2(\cdot\,;\boldsymbol{\theta})$, $E(T_1(\mathbf{y_i})) = f(\lambda_i)$ and $E(T_2(\mathbf{y_i})) = f^2(\lambda_i)$, it can be derived from equation (5)

$$\begin{aligned}
&\Sigma_i[\bar{x}_{i.} - T_1(\mathbf{y_i})] = 0 \\
&\Sigma_i[\bar{x}_{i.}^2 - T_2(\mathbf{y_i})] - IS_{xx}/J = 0\,.
\end{aligned} \tag{6}$$

Structural model. Similarly to the functional model, with $\mu_i = \frac{f(\lambda_i)-a}{b}$, the log likelihood function is

$$\begin{aligned} & l(x_{ij}, y_{ij}; \boldsymbol{\theta}, \mu_i) \\ & = \Sigma_i \Sigma_j log \int \frac{exp\{\frac{y_{ij} f(\lambda_i) - \psi_1(\lambda_i)}{\psi_2(\phi)} + c(y_{ij}, \phi) - \frac{(x_{ij}-\mu_i)^2}{2\sigma_\delta^2}\}}{\sqrt{2\pi}\sigma_\delta} p(\lambda_i) d\lambda_i \\ & = \Sigma_i \Sigma_j log \int p(x_{ij}, y_{ij}, \boldsymbol{\theta}, \mu_i) p(\lambda_i) d\lambda_i \,. \end{aligned}$$

It follows the score functions for $\boldsymbol{\theta}$ are

$$\begin{aligned} \frac{\partial}{\partial a} l &= -E[\Sigma_i \Sigma_j \frac{x_{ij}-\mu_i}{\sigma_\delta^2} | x_{ij}, y_{ij}] = 0 \\ \frac{\partial}{\partial b} l &= -E[\Sigma_i \Sigma_j \frac{(x_{ij}-\mu_i)\mu_i}{b\sigma_\delta^2} | x_{ij}, y_{ij}] = 0 \\ \frac{\partial}{\partial \sigma_\delta} l &= -IJ\frac{1}{\sigma_\delta} + E[\Sigma_i \Sigma_j \frac{(x_{ij}-\mu_i)^2}{b\sigma_\delta^3} | x_{ij}, y_{ij}] = 0 \\ \frac{\partial}{\partial \phi} l &= E[-\Sigma_i \Sigma_j \psi_2'(\phi) \frac{y_{ij} f(\lambda_i) - \psi_1(\lambda_i)}{\psi_2^2(\phi)} + \Sigma_i \Sigma_j \frac{\partial}{\partial \phi} c(y_{ij}, \phi) | x_{ij}, y_{ij}] = 0 \\ \mu_i &= \frac{f(\lambda_i) - a}{b} = 0 \,. \end{aligned} \tag{7}$$

Denote $\mathbf{y_i} = \{y_{ij} | j = 1, 2, ..., J\}$ and $S_{xx} = \Sigma_i \frac{(x_i j - \bar{x}_{i.})^2}{J-1}$. Assuming for some functions $T_1(\cdot\,; \boldsymbol{\theta})$,$T_2(\cdot\,; \boldsymbol{\theta})$, $E(T_1(\mathbf{y_i})) = f(\lambda_i)$ and $E(T_2(\mathbf{y_i})) = f^2(\lambda_i)$, it can be derived Based on law of total expectation from equation (5),

$$\begin{aligned} & \Sigma_i [\bar{x}_{i.} - T_1(\mathbf{y_i})] = 0 \\ & \Sigma_i [\bar{x}_{i.}^2 - T_2(\mathbf{y_i})] - IS_{xx}/J = 0 \,. \end{aligned} \tag{8}$$

3.4 Discussion

This section proposes a general idea to estimate intercept and slope in the platform comparison scenario when one of the platforms is not a continuous measurement. In practice, there are issues left for discussion. Firstly, for some distribution T_1 and T_2 do not have an explicit form and require a delta method which induces truncated error. Secondly, the efficiency of proposed estimator mostly depends on the efficiency of T_1 and T_2, therefore it is possible for some distributions that large sample size is required.

The latter issue is actually the major obstruction to applying the proposed method. However, in the real experiment, it is often feasible to acquire large number of subjects to measure. In next Chapter, we introduce a platform comparison with large number of measured genes with GEIV in a context of comparing Microarray and RNA-Seq. A brief introduction of those two measurement platforms can be find in Appendix 7.2.

4 Novel approach to compare Microarray and RNA-Seq with replicates with generalized linear EIV model

4.1 Background

Microarray and RNA-Seq are two major platforms for large scale transcriptomic profiling with the ability to measure thousands of genes simultaneously. Both platforms have been widely used for a long time in Biomedical research: Microarray is the traditional choice since mid 1990s and RNA-Seq emerges after 2005 as an attractive alternative (Marioni et al., 2008). Despite Microarray quantifies abundance with fluorescence intensity while RNA-Seq with fragment count, these two platforms both quantify transcript abundance in a sample. Increasingly popular in the recent years, the RNA-Seq technology has been used in a number of pathological studies such as prostate cancer(Ren et al., 2012), lung cancer(Beane et al., 2011) and breast cancer(Sinicropi et al., 2012), which are previously studied by Microarray also. As a consequence of such trend, whether measurements and analyses obtained from these two platforms are consistent with each other has become an intriguing and important issue.

As the newcomer, the RNA-Seq is naturally compared to the old favorite, the gene Microarray for measurement agreement. Some comparative researches have generated parallel data set that implemented both platforms on the same samples (Fu et al., 2009; Marioni et al., 2008; Mortazavi et al., 2008). In these studies, Pearson correlation and Spearman correlation are computed for quantifying the comparison and a strong positive correlation is observed between Microarray preprocessed intensity and RNA-Seq normalized count data in log scale (Xu et al., 2013). However, this comparison doesn't necessarily suggest consistency between measurements, as the consistency between platforms requires the constant bias is 0 and the proportional bias is 1 and a errors-in-variables (EIV) regression model is more appropriate for assessing the consistency (Linnet, 1993).

In (Xu et al., 2013), the linear EIV model is applied to the Microarray and

RNA-Seq to estimate those two types of biases and found a proportional bias between them. However, because linear EIV considers transformed RNA-Seq discrete count data value as a continuous random variable, the estimated biases only offers a rough description of the true biases. In order to construct an exact calibration model under more realistic assumptions, the linear EIV model between Microarray and RNA-Seq frame work is required to extend to GEIV. However, computing MLE of GEIV, or equivalently generalized linear model with measurement error, for platform comparison is also a challenging problem since the likelihood function is generally complex under the distribution assumption of Microarray and RNA-Seq and the score function has no analytical solutions. To bypass this difficulty, alternative estimate equations have been constructed for independent variable with normal measurement error such as conditional estimate equation, corrected score function and structural quasi score function (Kukush et al., 2004). The common prerequisites for those methods requires the measurement error variance to be known for Microarray or the number of replicates is large enough to replace true variance with the sample variance. However, to our knowledge, such large parallel data set is not available as most public data set only has 2-5 biological or technical replicates.

For genes with extremely low expressions, the background noise will affect the Microarray measurement and RNA-Seq measurement will constantly drop to 0 because of their technical nature. While for genes with extremely high expression, Microarrary suffers from the saturation. For these reasons, in this paper, we focus on genes that can be defined as 'commonly detectable' (Xu et al., 2013) based on usual filters (Mortazavi et al., 2008; Rau et al., 2013). The chapter proposes an approximate estimate algorithm of proposed generalized linear EIV without knowing the measurement error variance in independent variable. We further perform simulation verification on the proposed algorithm and then apply it to several available data sets to assess the consistency between Microarray and RNA-Seq.

4.2 EIV model for Microarray and RNA-Seq.

Let I denote the total number of genes, and J denote the total number of replicates. We set platform A to be the Microarray and platform B to be the RNA-Seq count, then ξ_{ij} and η_{ij} are the corresponding true expression profiles for gene i and replicate j on the two platfomrs. Specifically, ξ_{ij} is the theoretical fluorescence intensity for a certain abundance of transcripts while

η_{ij} is the theoretical fragment count of the same abundance of transcripts.

Transformation and true expression intensities. Although ξ_{ij} and η_{ij} are the true value of transcript abundance, recovering them from the observed data can be difficult because of the complexity of technology. Usually in further analysis the observed profile is transformed to bypass this issue. For this reason, it is defined the true expression intensity μ_{ij} and λ_{ij} as the expectation of the transformed observation for ξ_{ij} and η_{ij}.

For Microarray, after background correction, normalization (Huber et al., 2002). and log transformation (Durbin et al., 2002), we have $X_{ij} \sim N(\mu_{ij}, \sigma_i^2)$ (Smyth et al., 2004), where X_{ij} is the processed Microarray observations in log scale and σ_i is independent from μ_{ij}. In the other hand, we consider RNA-Seq to be a random selection process and the observed count can be modeled as a Poisson distribution whose mean is affected by the transcript abundance, gene length and library size (Salzman et al., 2011). Usually the observed count are adjusted by a library size factor L_j and a gene length factor l_i (Trapnell et al., 2012). Therefore, we have $\lambda_{ij} = E(Y_{ij})/L_j l_i = \eta_{ij}/L_j l_i$ and $Y_{ij} \sim Poisson(\lambda_{ij} L_j l_i)$.

As stated in Background, in general there is a strong Pearson correlation coefficient between these two platforms in log scale, therefore it is proposed the following generalized linear model to connect μ_{ij} and λ_{ij}:

$$\begin{aligned} log(\eta_{ij}) &= a + b\mu_{ij} + log(L_j l_i) \\ log(\lambda_{ij}) &= a + b\mu_{ij}\,. \end{aligned} \tag{9}$$

Types of replicates. In this chapter, two types of measurement replicates from a variance component view are defined, depending on whether there is additional source of variance in observed gene expression other than the measurement error itself. If measurement error is the only source of variance, for the replicates on the same gene, we define this type of replicates as constant replicates. It is the ideal type of technical replicates, where we can consider the difference in replicates is negligible comparing to the measurement error. If biological variation between replicates is expected, we define this type as random replicates which is also known as biological replicates . Different assumptions are made regarding which type of replicates available in datasets and subsequently the corresponding models of each type will be considered separately:

- **Constant replicates.** When we have reasons to assume that biological variation is much smaller than measurement error, the underlying true expressions can be considered as the same value among replicates, i.e. $\lambda_{ij} = \lambda_i$ and $\mu_{ij} = \mu_i$.

- **Random replicates.** If for the same gene, the biological variation is also a significant variance component, it is assumed the true gene expressions follow a distribution across replicates. For the reason of the widely used negative binomial distribution assumption on RNA-Seq data (Hardcastle and Kelly, 2010; Kvam et al., 2012; Robinson et al., 2010), the prior of true expression is assumed to be a gamma distribution (Γ) (Kvam et al., 2012). Consequently, the marginal distribution of RNA-Seq count follows is negative binomial given appropriate hyperparameters in prior. In summary, $\lambda_{ij} \sim \Gamma(\alpha_i, \beta_i)$ and $exp(a + b\mu_{ij}) \sim \Gamma(\alpha_i, \beta_i)$.

Generalized linear EIV model. The assumptions above can be summarized into following model:

$$\begin{aligned} Y_{ij} &\sim Possion(\lambda_{ij} N_{ij}) \\ X_{ij} &\sim N(\mu_{ij}, \sigma_i^2) \\ ln(\lambda_{ij}) &= a + b\mu_{ij} \\ \lambda_{ij} &= \lambda_i \ (constant\, replicates) \\ \lambda_{ij} &\sim \Gamma(\alpha_i, \beta_i) \ (random\, replicates) \\ N_{ij} &= lib_j \times length_i \,. \end{aligned} \tag{10}$$

The algorithm for estimating a and b for model (10) is introduced in following sections.

Assessing biases between Microarray and RNA-Seq. Given the nature of these two technologies, the quantified abundance μ_{ij} and $log\lambda_{ij}$ are essentially log ratios of the absolute abundance and their own reference levels. Therefore, with the estimated a and b for modeling a generalized linear relationship between Microarray and RNA-Seq , we need to further quantify the consistency by translating them to fixed and proportional biases by unifying the reference level. This can be done by applying percentile shift

normalization (Shwetha et al., 2013). Denote the new reference levels with Microarray log expression as μ_{ref} and RNA-Seq expression λ_{ref}. Again note that μ_{ij} and μ_{ref} are already in log scale, we have

$$\begin{aligned} log(\lambda_{ij}) &= a + b\mu_{ij} \\ log(\lambda_{ij}/\lambda_{ref}) &= [a - log(\lambda_{ref}) + b\mu_{ref}] + b(\mu_{ij} - \mu_{ref})\,. \end{aligned}$$

Therefore, in log scale, the fixed bias can be estimated by $a + log(\lambda_{ref}) - b\mu_{ref}$ and proportional bias can be estimated by b.

4.3 MLE of intercept and slope

In this section, it is shown that although the MLE estimators of proposed model is theoretically identifiable, their exact solution is not closed-form and difficult to compute, especially when the total gene number I is large.

4.3.1 Constant replicates model

Likelihood functions. For constant model in model (10), we can write its log likelihood function as,

$$\begin{aligned} l(y_{ij}, x_{ij}; a, b, \mu_i) &= \Sigma_i \Sigma_j y_{ij} log(\lambda_i N_{ij}) - \Sigma_i log(\sqrt{\frac{2\pi}{J}}\sigma_i) \\ &- \Sigma\Sigma_j \lambda_i N_{ij} - \Sigma_i \Sigma_j \frac{(x_{ij} - \mu_i)^2}{2\sigma_i^2} - \Sigma_i \Sigma_j log(y_{ij}!)\,. \end{aligned} \tag{11}$$

Number of genes: i=1 to I
Number of replicates: j=1 to J
Observed Microarray log intensity: x_{ij}
Observed RNA-Seq count: y_{ij}
Length factor library factor: N_{ij}(known)
μ_i and λ_i is the underlying true measures and $log\lambda_i = a + b\mu_i$.

MLE estimator. The partial derivatives of likelihood function (12) are:

$$\begin{aligned}
\frac{\partial l}{\partial a} &= \Sigma_i\Sigma_j y_{ij} - \Sigma_i\Sigma_j exp\{a + b\mu_i + log(N_{ij})\} = 0 \\
\frac{\partial l}{\partial b} &= \Sigma_i\Sigma_j y_{ij}\mu_i - \Sigma_i\Sigma_j \mu_i exp\{a + b\mu_i + log(N_{ij})\} = 0 \\
\frac{\partial l}{\partial \mu_i} &= \Sigma_j b y_{ij} - \Sigma_j b exp\{a + b\mu_i + log(N_{ij})\} - \Sigma_j \frac{x_{ij} - \mu_i}{\sigma_i^2} = 0 \quad (12) \\
\frac{\partial l}{\partial \sigma_i} &= J\sigma_i^2 - \Sigma_j(x_{ij} - \mu_i) = 0\,.
\end{aligned}$$

Theoretically MLE estimators can be obtained from solving the $2I + 2$ equations in equation (12), but note that the partial derivatives with respect to μ_i are transcendent equations, which makes it difficult to directly obtain the MLE.

4.3.2 Random replicates model

Likelihood functions. For random model in equation (10), the full log likelihood function is

$$\begin{aligned}
&l(y_{ij}, x_{ij}; a, b, \sigma_i^2, \alpha_i\beta_i, N_{ij}) = \quad (13) \\
&\Sigma_i\Sigma_j log \int p(y_{ij}|\lambda_{ij}; N_{ij}) p(x_{ij}|\lambda_{ij}; a, b, \sigma_i^2) p(\lambda_{ij}; \alpha_i, \beta_i) d\lambda_{ij},
\end{aligned}$$

where

$$\begin{aligned}
p(y_{ij}|\lambda_{ij}; N_{ij}) &= \frac{(N_{ij}\lambda_{ij})^{y_{ij}}}{y_{ij}!} e^{-N_{ij}\lambda_{ij}} \\
p(x_{ij}|\lambda_{ij}; a, b, \sigma_i^2) &= \frac{1}{\sqrt{2\pi}\sigma_i} exp\{-\frac{[x_{ij} - \mu_{ij}(\lambda_{ij})]^2}{2\sigma_i^2}\} \\
p(\lambda_{ij}; \alpha_i, \beta_i) &= \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} \lambda_{ij}^{\alpha_i - 1} e^{-\beta_i\lambda_{ij}} \\
\mu_{ij}(\lambda_{ij}) &= \frac{log\lambda_{ij} - a}{b}\,.
\end{aligned}$$

Number of genes: i=1 to I
Number of replicates: j=1 to J
Observed Microarray log intensity: x_{ij}
Observed RNA-Seq count: y_{ij}
Length factor library factor: N_{ij}(known)
μ_{ij} and λ_{ij} is the underlying true measures.

MLE estimator. The estimation of likelihood function (13) can be viewed as problem of estimating parameters of interest under empirical Bayesian frame work (Atchadé, 2011). The partial derivatives of a and b are:

$$\begin{aligned} \Sigma_i\Sigma_j \frac{x_{ij}}{IJ\sigma_i^2} - \Sigma_i\Sigma_j E_{\lambda_{ij}}(\frac{\mu_{ij}(\lambda_{ij})}{IJ\sigma_i^2}|x_{ij}, y_{ij}) &= 0 \qquad (14) \\ \Sigma_i\Sigma_j E(\frac{x_{ij}\mu_{ij}(\lambda_{ij})}{IJ\sigma_i^2}|x_{ij}, y_{ij}) - \Sigma_i\Sigma_j E_{\lambda_{ij}}(\frac{\mu_{ij}^2(\lambda_{ij})}{IJ\sigma_i^2}|x_{ij}, y_{ij}) &= 0 \end{aligned}$$

Directly solving equation system (14) requires computing 2I intractable integrals, which again makes it undesirable and difficult to directly solve MLE.

4.4 Estimation of constant replicate model

4.4.1 Transformation of MLE estimator of a, b

In platform comparison scenario, the only parameters of interest are the intercept and slope: a and b. Therefore, μ_i and σ_i are nuisance parameters. Even assume nuisance parameters are known, there is still not analytical solution for b. However, with some algebra, the MLE estimator in (12) can be simplified to

$$\begin{aligned} &\Sigma_i \frac{(\bar{x}_{i.} - \mu_i)}{\sigma_i^2} = 0 \\ &\Sigma_i \frac{\mu_i(\bar{x}_{i.} - \mu_i)}{\sigma_i^2} = 0 \\ &\Sigma_j (y_{ij} - exp\{a + b\mu_i + log(N_{ij})\}) = \Sigma_j \frac{x_{ij} - \mu_i}{b\sigma_i^2} \\ &\Sigma_j \frac{(x_{ij} - \mu_i)^2}{J} = \sigma_i^2 \,. \end{aligned}$$

Since Microarray and RNA-Seq are both measurement platforms of gene expression and $log(\lambda_i) = a + b\mu$, we can re-parametrize the first 2 equations above to

$$\begin{aligned} \Sigma_i \frac{(\bar{x}_{i.} - \mu_i)}{\sigma_i^2} &= 0 \\ \Sigma_i \frac{\mu_i\bar{x}_{i.} - \mu_i^2}{\sigma_i^2} &= 0 \,. \end{aligned}$$

From constant model in model (10), we have $\mu_i = \bar{x}_{i.} - \bar{\delta}_{i.}$, where $\bar{\delta}_{i.} = \Sigma_j \delta_{ij}$ and δ_{ij} is the measurement error of μ_i. That is,

$$\begin{aligned} \Sigma_i \frac{(\bar{x}_{i.} - \frac{log(\lambda_i)-a}{b})}{\sigma_i^2} &= 0 \\ \Sigma_i \frac{\bar{x}_{i.}^2 - \bar{\delta}_{i.}\bar{x}_{i.} - (\frac{log(\lambda_i)-a}{b})^2}{\sigma_i^2} &= 0\,. \end{aligned} \tag{15}$$

It can be rewritten as,

$$\begin{aligned} \hat{b} &= \frac{\Sigma_i \frac{log(\lambda_i)-\hat{a}}{\sigma_i^2}}{\Sigma_i \frac{\bar{x}_{i.}}{\sigma_i^2}} \\ \hat{b}^2 &= \frac{\Sigma_i \frac{(log(\lambda_i)-\hat{a})^2}{\sigma_i^2}}{\Sigma_i \frac{\bar{x}_{i.}^2 - \bar{\delta}_{i.}\bar{x}_{i.}}{\sigma_i^2}} \end{aligned} \tag{16}$$

Therefore, (15) is a second-degree polynomial equation system with respect to a and b. In this section, an estimator based on this equation system is presented, which is motivated the theory of corrected score function (Nakamura, 1990) and unbiased estimating equation (Cox, 1993) with unknown nuisance parameters.

4.4.2 Estimate without σ_i in denominator

From equation (15) we have,

$$\begin{aligned} \Sigma_i \frac{(\bar{x}_{i.} - \frac{log(\lambda_i)-a}{b})}{\sigma_i^2} &\sim \Sigma_i \frac{N(0, \frac{\sigma_i^2}{J})}{\Sigma_i \frac{1}{\sigma_i^2}} \\ \Sigma_i \frac{\bar{x}_{i.}^2 - \bar{\delta}_{i.}\bar{x}_{i.} - (\frac{log(\lambda_i)-a}{b})^2}{\sigma_i^2} &\sim \frac{\Sigma_i \frac{\mu_i^2}{\sigma_i^2} N(0, \frac{\sigma_i^2}{J\mu_i^2})}{\Sigma_i \frac{\mu_i^2}{\sigma_i^2}}\,. \end{aligned}$$

Therefore, the right-hand side of equation (15) can be presented as a weighted combination of zero mean normal random variables, where the weight is reversely proportional to the variance. Such combination guarantees the efficiency under normal assumptions. However, when I is large, we can relax

the efficiency requirement of the estimator by removing σ_i from the denominator a and the resulting estimator is still consistent but no longer efficient. Denote $\theta = \begin{bmatrix} a \\ b \end{bmatrix}$, the obtained estimating equations and solution are

$$f(\theta) = \begin{array}{c} \Sigma_i(\bar{x}_{i.} - \frac{log(\lambda_i)-a}{b}) \\ \Sigma_i[\bar{x}_{i.}^2 - \bar{\delta}_{i.}\bar{x}_{i.} - (\frac{log(\lambda_i)-a}{b})^2] \end{array} = 0 \tag{17}$$

and

$$\begin{aligned} \hat{a} &= \bar{log}(\lambda_.) - \hat{b}\bar{x}_{..} \\ \hat{b} &= \frac{\Sigma_i(log(\lambda_i) - \bar{log}(\lambda_.))^2}{\Sigma_i(\bar{x}_{i.} - \bar{x}_{..})^2 - \Sigma_i\bar{\delta}_{i.}\bar{x}_{i.}} \\ \bar{log}(\lambda_.) &= \frac{\Sigma_i log(\lambda_i)}{I} \\ \bar{x}_{..} &= \frac{\Sigma_i\Sigma_j x_{ij}}{IJ}\,. \end{aligned} \tag{18}$$

4.4.3 Approximately solving $f(\theta) = 0$

Lemma 4.1. $\sqrt{I}f(\theta) \xrightarrow{d} N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, V)$ *under reasonable regularity conditions, where V is a 2 by 2 matrix with finite elements.*

Proof. See Appendix 7.1.

Corollary 4.1. *Given* $\sqrt{I}f(\theta) \xrightarrow{d} N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, V)$ *and* $\hat{\theta}$ *is the solution of* $f(\theta) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, *we have* $\sqrt{I}(\hat{\theta} - \theta) \to N(0, U^{-1}VU'^{-1})$,*where*

$$U = \frac{1}{b}\begin{bmatrix} 1 & \frac{\sum_i \bar{x}_{i.}}{I} \\ \frac{2\sum_i \bar{x}_{i.}}{I} & \frac{2(\sum_i \bar{x}_{i.}^2 - \sum_i \sigma_i^2/J)}{I} \end{bmatrix}.$$

Proof. By Taylor expansion

$$0 = f(\hat{\theta}) = f(\theta) + \nabla f(\theta)(\hat{\theta} - \theta) + o(\sqrt{I}),$$

therefore $\sqrt{I}(\hat{\theta}-\theta) \xrightarrow{p} [\nabla f(\theta)]^{-1}\sqrt{I}f(\theta)$. From law of large number (LLN),

$$\nabla f(\theta) = \frac{1}{b}\begin{bmatrix} 1 & \frac{\sum_i \mu_i}{I} \\ \frac{2\sum_i \mu_i}{I} & \frac{2\sum \mu_i^2}{I} \end{bmatrix}$$
$$\lim U \to \nabla f(\theta)\,.$$

Therefore $\sqrt{I}(\hat{\theta}-\theta) \to N(0, U^{-1}VU'^{-1})$.

Lemma 4.2. *Under regularity conditions, when $y_{i.} = \Sigma_j y_{ij} > 0$ and $I \to \infty$. Denote $N_{i.} = \Sigma_j N_{ij}$,*

$$\Sigma_i[log(\frac{y_{i.}}{N_{i.}}) + \frac{1}{2(y_{i.}+1)}] \approx \Sigma_i log(\lambda_i)$$
$$\Sigma_i\Sigma_j \frac{(x_{ij}-\bar{x}_{i.})^2}{J(J-1)} \to \Sigma_i \bar{\delta}_{i.}\bar{x}_{i.}$$
$$\Sigma_i[log(\frac{y_{i.}}{N_{i.}}) + \frac{1}{2(y_{i.}+1)}]^2 - \Sigma_i \frac{1}{y_{i.}+1} \approx \Sigma_i log^2(\lambda_i)\,.$$

Proof. See Appendix 7.1.

Thereom 1. *Under regularity conditions, let*

$$l_{.} = log(\frac{y_{i.}}{N_{i.}}) + \frac{1}{2(y_{i.}+1)}$$
$$\bar{l}_{.} = \frac{\Sigma_i l_i}{I}$$
$$\bar{x}_{i.} = \frac{\Sigma_j x_{ij}}{J}$$
$$\bar{x}_{..} = \frac{\Sigma_i\Sigma_j x_{ij}}{IJ},$$

the following estimator is approximate to a consistent estimator

$$\begin{aligned} \hat{a} &= \bar{l}_{.} - \hat{b}\bar{x}_{..} \\ \hat{b} &= \frac{\Sigma_i(l_i - \bar{l}_{.})^2 - \Sigma_i \frac{1}{y_{i.}+1}}{\Sigma_i(\bar{x}_{i.} - \bar{x}_{..})^2 - \Sigma_i\Sigma_j \frac{(x_{ij}-\bar{x}_{i.})^2}{J(J-1)}}\,. \end{aligned} \tag{19}$$

Proof. Equation (19) can be rewritten as

$$\begin{aligned} \Sigma_i(\bar{x}_{i.} - \frac{l_i - a}{b}) &= 0 \\ \Sigma_i(\bar{x}_{i.})^2 - \Sigma_i\Sigma_j \frac{(x_{ij} - \bar{x}_{i.})^2}{J(J-1)} - \frac{\Sigma_i(l_i - a)^2 - \Sigma_i \frac{1}{y_{i.}+1}}{b^2} &= 0\,. \end{aligned}$$

From **Lemma 4.2**, it can be shown the above equation system approximate to equation (17). Therefore, the solution is approximate to the solution of (17) which is shown to be a consistent estimator in **Lemma 4.1**.

4.5 Estimation of random replicate model

4.5.1 Unbiased estimating equation

Lemma 4.3. *Under reasonable regularity conditions.*

$$\begin{aligned} \Sigma_i\Sigma_j E_{\lambda_{ij}}(\frac{\mu_{ij}(\lambda_{ij})}{IJ\sigma_i^2}|x_{ij}, y_{ij}) &\xrightarrow{p} \Sigma_i\Sigma_j E_{\lambda_{ij}}(\frac{\mu_{ij}(\lambda_{ij})}{IJ\sigma_i^2}) \\ \Sigma_i\Sigma_j E_{\lambda_{ij}}(\frac{\mu_{ij}^2(\lambda_{ij})}{IJ\sigma_i^2}|x_{ij}, y_{ij}) &\xrightarrow{p} \Sigma_i\Sigma_j E_{\lambda_{ij}}(\frac{\mu_{ij}^2(\lambda_{ij})}{IJ\sigma_i^2})\,. \end{aligned}$$

Proof. see Appendix 7.1.

From an empirical point of view, we can consider the regularity conditions are reasonable, given that the variance of observed value for subjects is unlikely to be comparable to I. Even if some transcripts have extremely large variance, we can filter it out in an analysis because it provides little information.

Lemma 4.4. *Under reasonable regularity conditions,*

$$\Sigma_i\Sigma_j E_{\lambda_{ij}}(\frac{x_{ij}\mu_{ij}(\lambda_{ij})}{IJ\sigma_i^2}|x_{ij}, y_{ij}) \xrightarrow{p} \Sigma_i\Sigma_j \frac{x_{ij}^2}{IJ\sigma_i^2} - 1\,.$$

Proof. see Appendix 7.1.

Therefore (14) can be approximated by

$$\begin{aligned} \Sigma_i\Sigma_j \frac{x_{ij}}{IJ\sigma_i^2} - \Sigma_i\Sigma_j \frac{\psi(\alpha_i) - log\beta_i - a}{IJb\sigma_i^2} &= 0 \qquad (20) \\ \Sigma_i\Sigma_j \frac{x_{ij}^2}{IJ\sigma_i^2} - 1 - \Sigma_i\Sigma_j(\frac{\psi_1(\alpha_i) + [\psi(\alpha_i) - log\beta_i - a]^2}{IJb^2\sigma_i^2}) &= 0\,. \end{aligned}$$

Similar to constant model, we can remove σ_i^2 from the denominator:

$$\begin{aligned} \Sigma_i\Sigma_j \frac{x_{ij}}{IJ} - \Sigma_i\Sigma_j \frac{\psi(\alpha_i) - log\beta_i - a}{IJb} &= 0 \quad (21) \\ \Sigma_i\Sigma_j \frac{x_{ij}^2}{IJ} - \Sigma_i\Sigma_j \frac{x_{ij}\delta_{ij}}{IJ} - \Sigma_i\Sigma_j (\frac{\psi_1(\alpha_i) + [\psi(\alpha_i) - log\beta_i - a]^2}{IJb^2}) &= 0\,. \end{aligned}$$

Where ψ is digamma function and ψ_1 is trigamma function. Assuming the nuisance parameters σ_i,α_i and β_i are fixed, define $\theta = \begin{bmatrix} a \\ b \end{bmatrix}$, from (21) we can construct an estimating function as

$$h(\theta) = \begin{matrix} \Sigma_i\Sigma_j \frac{x_{ij}}{IJ} - \Sigma_i\Sigma_j \frac{\psi(\alpha_i) - log\beta_i - a}{IJb} \\ \Sigma_i\Sigma_j \frac{x_{ij}^2}{IJ} - \Sigma_i\Sigma_j \frac{x_{ij}\delta_{ij}}{IJ} - \Sigma_i\Sigma_j (\frac{\psi_1(\alpha_i) + [\psi(\alpha_i) - log\beta_i - a]^2}{IJb^2}) \end{matrix}.$$

And the estimator is the solution of $h(\theta) = 0$.

4.5.2 Approximately solving $h(\theta) = 0$

It follows that $E(h(\theta)) = 0$, thus $h(\theta) = 0$ is an UEE. Extending the asymptotic theory of unbiased estimation equation (Cox, 1993), we have

Lemma 4.5. *$\sqrt{IJ}h(\theta) \xrightarrow{d} N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, V)$ under reasonable regularity conditions, where V is 2 by 2 matrix with finite values.*

Proof. see Appendix 7.1.

Corollary 4.2. *Given $\sqrt{IJ}h(\theta) \xrightarrow{d} N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, V)$ and $\hat{\theta}$ is the solution of $h(\theta) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, we have $\sqrt{IJ}(\hat{\theta} - \theta) \to N(0, U^{-1}VU'^{-1})$, where*

$$U = \frac{1}{b}\begin{bmatrix} 1 & \frac{\sum_i \sum_j x_{ij}}{IJ} \\ \frac{2\sum_i \sum_j x_{ij}}{IJ} & \frac{2(\sum_i \sum_j x_{ij}^2 - \sum_i \sum_j \sigma_{ij}^2)}{IJ} \end{bmatrix}.$$

Proof. By Taylor expansion

$$0 = h(\hat{\theta}) = h(\theta) + \nabla h(\theta)(\hat{\theta} - \theta) + o(\sqrt{IJ}),$$

therefore $\sqrt{IJ}(\hat{\theta}-\theta) \xrightarrow{p} [\nabla h(\theta)]^{-1}\sqrt{IJ}h(\theta)$. By LLN,

$$\nabla h(\theta) = \frac{1}{b}\begin{bmatrix} 1 & \frac{\sum_i \sum_j E(\mu_{ij})}{IJ} \\ \frac{2\sum_i \sum_j E(\mu_{ij})}{IJ} & \frac{2\sum \sum_j E(\mu_{ij}^2)}{IJ} \end{bmatrix}$$
$$\lim U \to \nabla h(\theta)\,.$$

Consequently, $\sqrt{IJ}(\hat{\theta}-\theta) \to N(0, U^{-1}VU'^{-1})$. The solution of $h(\theta) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ is not an efficient estimator.However, it is consistent with a reasonable efficiency, which makes it a practical estimator for a and b when I is large. It can be shown that

Lemma 4.6. *Under regularity conditions, assuming minimal coefficient of variation of λ_{ij} is greater than 1. Let $l_{ij} = log(\frac{y_{ij}}{N_{ij}}) + \frac{1}{2(y_{ij}+1)}$,*

$$\begin{aligned}
&\Sigma_i\Sigma_j l_{ij} \approx \Sigma_i J(\psi(\alpha_i) - log(\beta_i)) \\
&\Sigma_i \frac{(x_{ij}-\bar{x}_{i.})^2}{(J-1)} \to \Sigma_i(\sigma_i^2 + \frac{\psi_i(\alpha_i)}{b^2}) \\
&\Sigma_i\Sigma_j(\frac{l_{ij}-a}{b})^2 + \Sigma_i\Sigma_j \frac{J(x_{ij}-\bar{x}_{i.})^2}{J-1} - \Sigma_i\Sigma_j \frac{J(l_{ij}-\bar{l}_{i.})^2}{b^2(J-1)} \\
&\approx \Sigma_i\Sigma_j \frac{[\psi(\alpha_i) - log(\beta_i) - a]^2 + \psi_1(\alpha_i)}{b^2} + \Sigma_i\Sigma_j\sigma_i^2\,.
\end{aligned}$$

Proof. see Appendix 7.1.

Consequently, the estimating equation can be derived.

Thereom 2. *Under regularity conditions, let*

$$
\begin{aligned}
l_{ij} &= log(\frac{y_{ij}}{N_{ij}}) + \frac{1}{2(y_{ij}+1)} \\
\bar{l}_{i.} &= \Sigma_j \frac{l_{ij}}{J} \\
\bar{l}_{..} &= \Sigma_i \Sigma_j \frac{l_{ij}}{J} \\
s_i^2(l) &= \Sigma_j \frac{(l_{ij} - \bar{l}_{i.})^2}{J-1} \\
s_i^2(x) &= \Sigma_j \frac{(x_{ij} - \bar{x}_{i.})^2}{J-1} \\
\bar{x}_{i.} &= \Sigma_j \frac{x_{ij}}{J} \\
\bar{x}_{..} &= \frac{\Sigma_i \Sigma_j x_{ij}}{IJ} \,.
\end{aligned}
$$

The following estimator is approximate to a consistent estimator:

$$
\begin{aligned}
\hat{a} &= \bar{l}_{..} - \hat{b}\bar{x}_{..} \\
\hat{b} &= \frac{\Sigma_i \Sigma_j (l_{ij} - \bar{l}_{..})^2 - \Sigma_i \Sigma_j s_i^2(l)}{\Sigma_i \Sigma_j (x_{ij} - \bar{x}_{..})^2 - \Sigma_i \Sigma_j s_i^2(x)} \,.
\end{aligned} \tag{22}
$$

Proof. Similar to Theorem 1 with Lemma 4.5 and 4.6.

4.6 Filter of low count data in RNA-Seq

From empirical point of view in RNA-Seq platform, the genes with low aligned count are considered as low-informative (Anders et al., 2013). In additional, based on the model (10), the coefficient of variation of count distribution is $\frac{1}{N_{ij}\lambda ij}$, which increases rapidly when the expectation of count $N_{ij}\lambda ij$ decreases to 0. Especially the delta method is used, the proposed estimators are largely affected by low count data and $log(0) = -\infty$. Therefore, it is desirable to filter out the low count data before analysis. Current methods to set up cutoff of the filter include empirical cutoff (Mortazavi et al., 2008) and data-driven cutoff (Rau et al., 2013). In this dissertation, we adopted a data-driven cutoff on observed aligned counts of gene as following:

1. After removing the genes with 0 total counts from all replicates, the number of remaining genes, denoting as n.

2. Tentative cutoffs are set at the lowest 0.5%,1%,1.5%...to 5% mean observed expression intensity value $\frac{1}{J}\Sigma_j \frac{y_{ij}}{N_{ij}}$. Given a cutoff s, let n_1^s denote the number of genes that all replicates pass s and n_2^s denote the number of genes at least one replicate pass s.

3. The final cutoff is $S = argmin_s\{\frac{n_1^s}{n_2^s}\}$, and for a gene i, if $max_j l_{ij} < S$, it is filtered out.

4. For random model, the remaining zeros for gene i and replicate j will be excluded in estimators (see Appendix 7.3).

4.7 Preprocessing variation and Modeling error

Before taking the gene expression measures, both Microarray and RNA-Seq have a preprocessing stage for the reverse transcription of RNA molecules and amplification of resulting cDNA. However, studies (Hyman et al., 2002; Ståhlberg et al., 2004) has revealed both procedures in the preprocessing add additional variation in the gene expression profiles. Consequently, even within each platform, a batch effect has been observed (Chen et al., 2011; Sun et al., 2013). After taking the preprocessing variation into consideration, the platforms' true expressions are only 'true' with respect to measurement error, hence they are not no longer expected to distribute right on a straight line. Therefore in this section, we further extend the GEIV model (10) under a normal assumption of preprocessing error, which models the pattern of the true expressions as a ellipse formed by a bivariate normal distribution (see Figure 3).

For the constant replicates model (see Figure 4), we assume that given the real RNA abundance ξ_i for a gene in a sample, it is affected by the error $\Delta_{i,array}$ and $\Delta_{i,seq}$ in the preprocessing for Microarray and RNA-Seq respectively. Note that this error is multiplicative on the raw expressions but becomes additive in the log scale. The resulting true expressions for both platforms are $\xi_{i,array}$ and $\xi_{i,seq}$. From the platform comparison model (10), we have $\mu_i = \xi_{i,array}$ and $log(\lambda_i) = a + b\xi_{i,seq}$. Assuming $\Delta_{i,array}$, $\Delta_{i,seq} \sim N(0, \tau_\Delta^2)$ and $\xi_i \sim N(\upsilon, \tau_\xi^2)$. Subsequently, the proposed constant replicates

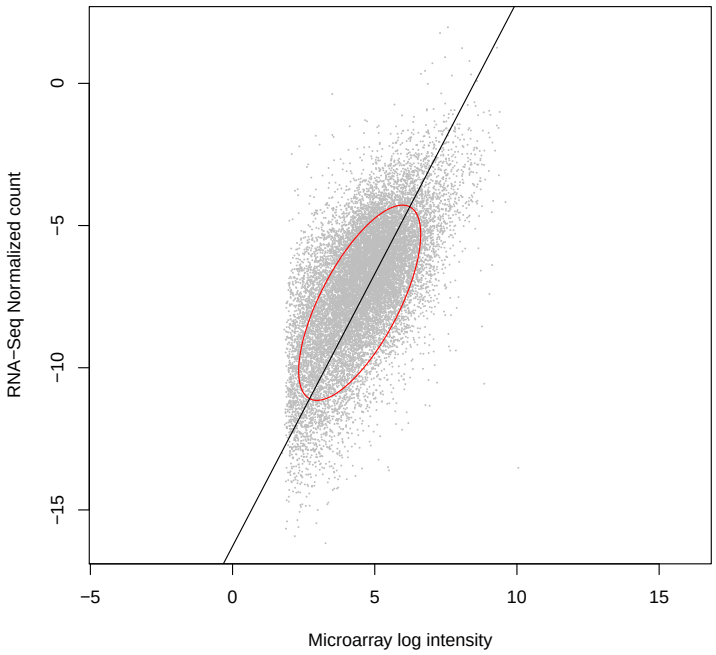

Figure 3: Ellipse of bivariate normal distribution

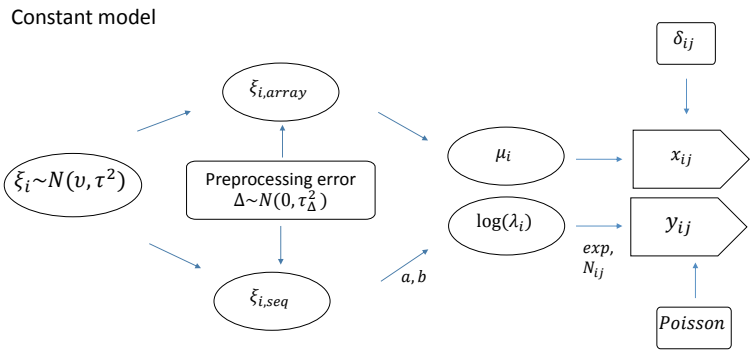

Figure 4: Constant replicates model with preprocessing error

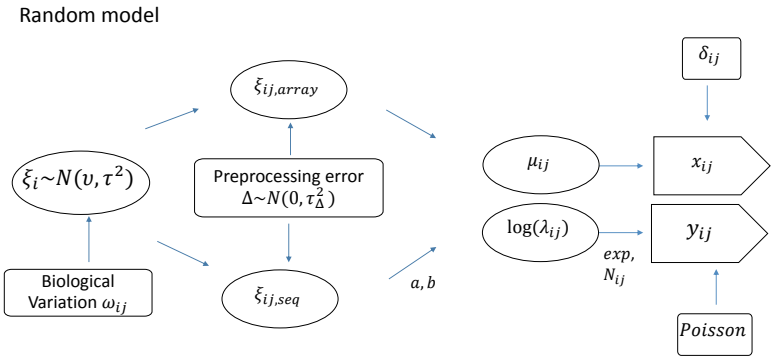

Figure 5: Random replicates model with preprocessing error

model with preprocessing error is

$$\begin{aligned} Y_{ij} &\sim Possion(\lambda_i N_{ij}) \\ X_{ij} &\sim N(\mu_i, \sigma_i^2) \\ \begin{pmatrix} \mu_i \\ log(\lambda_i) \end{pmatrix} &\sim N\left(\begin{pmatrix} \upsilon \\ a + b\upsilon \end{pmatrix}, \begin{pmatrix} \tau_\xi^2 + \tau_\Delta^2 & b\tau_\xi^2 \\ b\tau_\xi^2 & b^2(\tau_\xi^2 + \tau_\Delta^2) \end{pmatrix}\right). \end{aligned} \tag{23}$$

For the random replicates model (see Figure 5), although there is a zero-mean biological variation term ω_{ij} in the abundance of a gene among replicates, the condition distribution of $(\mu_{ij}, log(\lambda_{ij}))^T$ given ω_{ij} is still a bivariate normal distribution. Note that $\lambda_{ij} \sim \Gamma(\alpha_i, \beta_i)$ because of ω_{ij}. With a more straightforward parameterization of the gamma prior, we have

$$\begin{aligned} Y_{ij} &\sim Possion(\lambda_{ij} N_{ij}) \\ X_{ij} &\sim N(\mu_{ij}, \sigma_i^2) \\ \begin{pmatrix} \mu_{ij} \\ log(\lambda_{ij}) \end{pmatrix} | \omega_{ij} &\sim N\left(\begin{pmatrix} \upsilon + \omega_{ij} \\ a + b(\upsilon + \omega_{ij}) \end{pmatrix}, \begin{pmatrix} \tau_\xi^2 + \tau_\Delta^2 & b\tau_\xi^2 \\ b\tau_\xi^2 & b^2(\tau_\xi^2 + \tau_\Delta^2) \end{pmatrix}\right) \\ \lambda_{ij} &\sim \Gamma(\alpha_i, \frac{\alpha_i}{exp\{a + b\xi_{i,seq}\}}) \\ \mu_{ij} &= \xi_{i,array} + \frac{log(\lambda_{ij}) - log(\xi_{i,seq})}{b}. \end{aligned} \tag{24}$$

Although the marginal distribution of $(\mu_{ij}, log(\lambda_{ij}))^T$ is no long bivariate normal, the expectations of replicates across genes are still distributed on an ellipse because $E(\omega_{ij})$ is assumed to be 0.

In order to save notations, we denote $\sigma_x^2 = var(\mu_i)$, $\sigma_y^2 = var(log(\lambda_i))$ and $\rho = cor(\mu_i, log(\lambda_i))$ for constant replicates model (23). And $\sigma_x^2 = var(\mu_{ij})$, $\sigma_y^2 = var(log(\lambda_{ij}))$ and $\rho = cor(\mu_{ij}, log(\lambda_{ij}))$ for random replicates model (24). Our goal is to estimate the covariance matrix $\Lambda = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$. The reason to estimate this matrix is that although the estimators of a and b remain the same with preprocessing error, the line $y = a + bx$ is no longer the optimal line to describe the pattern of true intensities. Instead, it is described by the ellipse corresponding to Λ and its major axis which has the same slope with the first principal component, which is denoted as $y = PC_a + PC_b x$.

Lemma 4.7. *Under model* (23), *let* $l_i = log(\frac{y_{i.}}{N_{i.}}) + \frac{1}{2(y_{i.}+1)}, \bar{l}_{.} = \frac{\Sigma_j l_i}{J}$, $y_{i.} = \Sigma_j y_{ij} > 0$ *and* $N_{i.} = \Sigma_j N_{ij}$

$$\begin{aligned}
&\Sigma_i \bar{x}_{i.}^2 - \Sigma_i \Sigma_j \frac{(x_{ij} - \bar{x}_{i.})^2}{J(J-1)} - I\bar{x}_{..}^2 \approx I\sigma_x^2 \\
&\Sigma_i (l_i)^2 - \Sigma_i \frac{1}{y_{i.}+1} - I\bar{l}_{.}^{\,2} \approx \ I\sigma_y^2 \\
&\Sigma_i \bar{x}_{i.} l_i - I\bar{x}_{..}\bar{l}_{.} \approx \ I\rho\sigma_x\sigma_y \,.
\end{aligned}$$

Proof. see Appendix 7.1.

Lemma 4.8. *Under model* (23), *let* $l_{ij} = log(\frac{y_{ij}}{N_{ij}}) + \frac{1}{2(y_{ij}+1)}, \bar{l}_{i.} = \frac{\Sigma_j l_{ij}}{J}, \bar{l}_{..} = \frac{\Sigma_i \Sigma_j l_{ij}}{IJ}, s_i^2(l) = \Sigma_j \frac{(l_{ij} - \bar{l}_{i.})^2}{J-1}$ *and* $s_i^2(x) = \Sigma_j \frac{(x_{ij} - \bar{x}_{i.})^2}{J-1}$

$$\begin{aligned}
&\Sigma_i \Sigma_j x_{ij}^2 - \Sigma_i \Sigma_j s_i^2(x) - IJ\bar{x}_{..}^2 \approx IJ\sigma_x^2 \\
&\Sigma_i \Sigma_j (l_{ij})^2 - \Sigma_i \Sigma_j s_i^2(l) - IJ\bar{l}_{..}^{\,2} \approx \ IJ\sigma_y^2 \\
&\Sigma_i \Sigma_j x_{ij} l_{ij} - IJ\bar{x}_{..}\bar{l}_{..} \approx \ IJ\rho\sigma_x\sigma_y \,.
\end{aligned}$$

Proof. see Appendix 7.1.

Thereom 3. *Given model* (23) *and* (24). *a and b can be estimated respectively from matrices:*

$$\frac{1}{I}\begin{bmatrix} \Sigma_i \bar{x}_{i.}^2 - \Sigma_i \Sigma_j \frac{(x_{ij} - \bar{x}_{i.})^2}{J(J-1)} - I\bar{x}_{..}^2 & \Sigma_i \bar{x}_{i.} l_i - I\bar{x}_{..}\bar{l}_{.} \\ \Sigma_i \bar{x}_{i.} l_i - I\bar{x}_{..}\bar{l}_{.} & \Sigma_i (l_i)^2 - \Sigma_i \frac{1}{y_{i.}+1} - I\bar{l}_{.}^{\,2} \end{bmatrix} \tag{25}$$

and

$$\frac{1}{IJ}\begin{bmatrix} \Sigma_i\Sigma_j x_{ij}^2 - \Sigma_i\Sigma_j s_i^2(x) - IJ\bar{x}_{\cdot\cdot}^2 & \Sigma_i\Sigma_j x_{ij}l_{ij} - IJ\bar{x}_{\cdot\cdot}\bar{l}_{\cdot\cdot} \\ \Sigma_i\Sigma_j x_{ij}l_{ij} - IJ\bar{x}_{\cdot\cdot}\bar{l}_{\cdot\cdot} & \Sigma_i\Sigma_j (l_{ij})^2 - \Sigma_i\Sigma_j s_i^2(l) - IJ\bar{l_{\cdot\cdot}}^2 \end{bmatrix}. \quad (26)$$

Proof. It is easy to show that the above matrices approximate Λ under model (23) and (24) with Lemma 4.7 and 4.8. It follows that the major axis' slope is the same with the corresponding direction of first principle component of Λ.

4.8 Simulation study

4.8.1 Simulation 1

Constant model simulation without preprocessing error. In this simulation study, we first generated true RNA-Seq gene expression intensity using the same strategy proposed in (Xu et al., 2013) with a human T-cell dataset (Zhao et al., 2014). To be more specific, reads per kilobase of exon per million reads mapped (RPKM) values for 21498 genes in dataset were extracted as well as the gene length. A log-normal distribution was fitted with the RPKM values using MLE method and then a random sample with same number of genes was generated from this distribution as the true expression intensities λ_i in RPKM scale. In addition, to retain the structure of real RNA-Seq dataset, simulated true intensities are mapped to real genes with a turbulence (see Appendix 7.3) and then were assigned with the real genes' lengths. The mean of Microarray true intensities in log scale were set to $\mu_i = \frac{log(\lambda_i)-a}{b}$ with given a and b. Estimators of a and b were computed at different combination of σ_i^2 and library size factors. The range of those parameters were set as following based on real dataset in (Marioni et al., 2008)

$$\sigma_i \sim logN(log(\mu_\sigma), 0.25), \mu_\sigma \sim U(0.3, 0.8)$$
$$lib_j \sim m * N(1, 0.01), m \sim 10^{U(0,1)}$$
$$b \sim U(0.5, 2), a \sim N(E[log(\lambda_i)] - 5b, 0.04b^2)$$

and the generated genes were filtered according to Section 4.6.

A total of 50 sets of parameters were generated. With each set of parameters, number of replicates J was chosen at 3 levels: $(3, 5, 10)$ and 10 runs were repeated for each level for J. In total, 1500 runs of simulation were generated. The estimates are shown in Table 1 and Figure 6.

Simulation result. Although enlarging the sample size of replicates displays a minor reduction on bias and standard deviation of estimates, with all J=3,5 and 10, the small variation in estimates shows the convergence of proposed estimator does not require large number of replicates. However, the estimator is approximately consistent, therefore a small bias is observed. This bias is mostly caused by the 5%-15% genes with zero counts combining all replicates has to be filtered out due to the logarithm. Therefore, for lowly expressed genes, the information provided by data is complete and distribution is skewed. By increasing the library size factor, the number of zero-count genes decreases, hence the estimation bias reduces.

Parameters					Estimate of intercept		Estimate of slope	
$E(\sigma_i^2)$	a	b	J	Lib factor	Mean	SD	Mean	SD
0.404	-5.877	1.046	3	4.689	-5.594	0.017	1.014	0.002
0.404	-5.877	1.046	5	4.689	-5.649	0.017	1.02	0.002
0.404	-5.877	1.046	10	4.689	-5.757	0.007	1.032	0.001
0.407	-9.543	1.812	3	2.414	-9.086	0.039	1.749	0.006
0.407	-9.543	1.812	5	2.414	-9.14	0.021	1.756	0.003
0.407	-9.543	1.812	10	2.414	-9.251	0.017	1.771	0.003
0.724	-4.511	0.725	3	2.032	-4.225	0.015	0.7	0.002
0.724	-4.511	0.725	5	2.032	-4.263	0.011	0.703	0.001
0.724	-4.511	0.725	10	2.032	-4.316	0.008	0.708	0.001
0.656	-6.582	1.238	3	1.458	-6.187	0.024	1.191	0.003
0.656	-6.582	1.238	5	1.458	-6.213	0.021	1.193	0.003
0.656	-6.582	1.238	10	1.458	-6.295	0.014	1.203	0.002
0.352	-3.19	0.505	3	1.459	-2.92	0.005	0.485	0
0.352	-3.19	0.505	5	1.459	-2.946	0.008	0.487	0.001
0.352	-3.19	0.505	10	1.459	-2.998	0.008	0.49	0.001

Table 1: Result of Simulation 1(truncated)

Comparison with other estimator. The proposed estimator was also compared with others listed in Table 2 and the result is shown in Figure 7.

Estimator	Description
Naïve x	The GLM fit considering Microarray is exactly measured
Naïve y	The GLM fit considering RNA-Seq is exactly measured
GEIV	Proposed method
EIV	Structural linear EIV in Section 2.2
GMR	Geometric Mean of two Naïve estimators

Table 2: List of compared estimators in Simulation 1 and 2

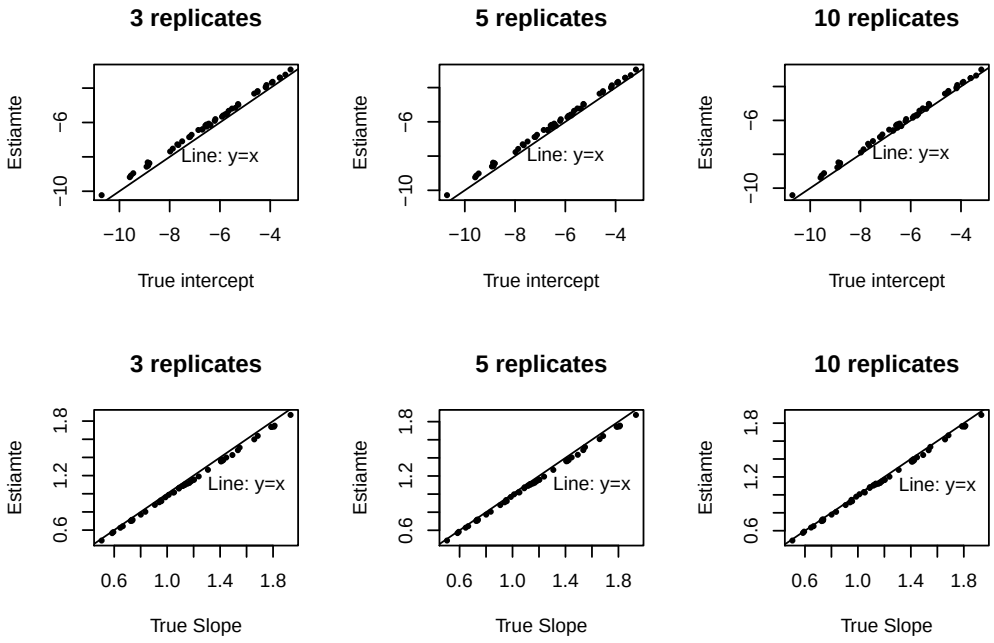

Figure 6: Simulation 1

Plotted true value versus average estimate

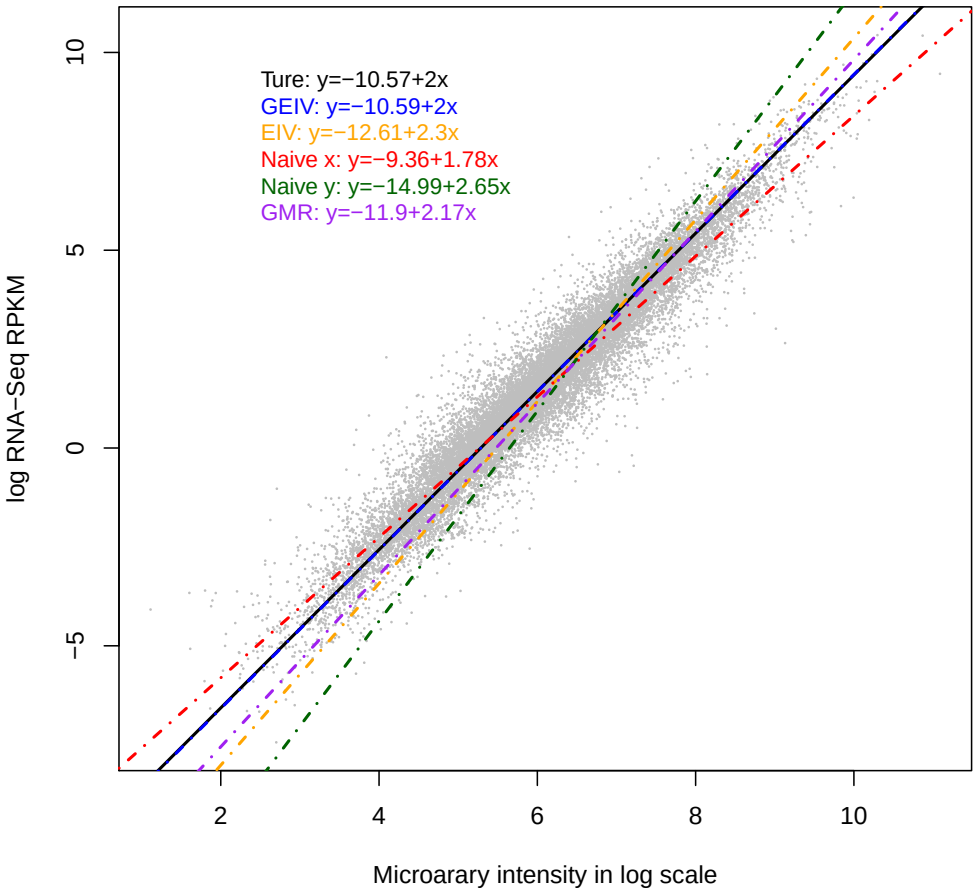

Figure 7: Comparison of estimators Simulation 1

4.8.2 Simulation 2

Constant model simulation with preprocessing error. Given a set of a and b, we first generated gene lengths and λ_i with the same method of Simulation 1 and then set $\tau_\xi^2 = var(log(\lambda_i))$ as well as $\upsilon = E(log(\lambda_i))$. Given a correlation coefficient ρ, μ_i was generated from conditional distribution $N\left(\upsilon + \frac{\rho}{b}(log(\lambda_i) - \upsilon), \frac{\tau_\xi^2}{b}(1-\rho^2)\right)$ as the true intestines for Microarray. Different from modeling without error, the desired line here is no longer $y = a + bx$, but the major axis of the ellipse generated by above normal distribution: $y = PC_a + PC_b x$. ρ was randomly generated from $U[0.65 - 0.9]$ based on current parallel Microarray and RNA-Seq datasets. Other parameters were the same with Simulation 1. Estimators of a and b were computed at different combination of parameters. A total of 50 sets of parameters were generated. With each set of parameters, the number of replicates J wass chosen at 3 levels: $(3, 5, 10)$ and 10 runs are repeated for each level for J. In total 1500 runs of simulation were generated. The generated genes were filtered according to Section 4.6 and the estimates are shown in Table 3 and Figure 8.

Parameters								Esimates of a and b				Estimates of ellipse major axis			
$E(\sigma_i^2)$	a	b	J	E(lib)	ρ	PC_a	PC_b	$E(a)$	$sd(a)$	$E(b)$	sd(b)	$E(PC_a)$	$sd(PC_a)$	$E(PC_b)$	$sd(PC_b)$
0.528	-9.071	1.701	3	4.152	0.801	-10.4	1.913	-8.726	0.048	1.653	0.007	-10.08	0.063	1.868	0.01
0.528	-9.071	1.701	5	4.152	0.801	-10.4	1.913	-8.811	0.062	1.663	0.01	-10.16	0.101	1.877	0.016
0.528	-9.071	1.701	10	4.152	0.801	-10.4	1.913	-8.904	0.046	1.677	0.008	-10.24	0.057	1.89	0.009
0.602	-8.85	1.807	3	1.389	0.739	-10.87	2.159	-8.106	0.05	1.698	0.009	-10.13	0.085	2.045	0.015
0.602	-8.85	1.807	5	1.389	0.739	-10.87	2.159	-8.3	0.05	1.726	0.008	-10.34	0.094	2.076	0.015
0.602	-8.85	1.807	10	1.389	0.739	-10.87	2.159	-8.486	0.038	1.753	0.007	-10.54	0.066	2.107	0.013
0.588	-9.306	1.783	3	1.161	0.729	-11.5	2.144	-8.471	0.055	1.668	0.009	-10.66	0.101	2.021	0.017
0.588	-9.306	1.783	5	1.161	0.729	-11.5	2.144	-8.598	0.061	1.683	0.01	-10.77	0.087	2.035	0.014
0.588	-9.306	1.783	10	1.161	0.729	-11.5	2.144	-8.838	0.036	1.717	0.006	-11.03	0.04	2.074	0.006
0.518	-5.092	0.927	3	1.315	0.672	-4.852	0.893	-4.491	0.03	0.863	0.003	-3.992	0.038	0.795	0.004
0.518	-5.092	0.927	5	1.315	0.672	-4.852	0.893	-4.645	0.035	0.878	0.005	-4.214	0.05	0.819	0.007
0.518	-5.092	0.927	10	1.315	0.672	-4.852	0.893	-4.768	0.052	0.891	0.006	-4.395	0.069	0.84	0.009
0.78	-4.626	0.767	3	3.948	0.861	-4.374	0.736	-4.415	0.021	0.747	0.002	-4.118	0.022	0.71	0.003
0.78	-4.626	0.767	5	3.948	0.861	-4.374	0.736	-4.468	0.023	0.752	0.003	-4.181	0.025	0.716	0.003
0.78	-4.626	0.767	10	3.948	0.861	-4.374	0.736	-4.519	0.019	0.757	0.003	-4.246	0.021	0.723	0.003

Table 3: Result of Simulation 2 (truncated)

Simulation result. Same to shown Simulation 1, enlarging the sample size of replicates displays a minor reduction on bias and standard deviation of estimates, estimator in equation (25) shows convergence with J=3,5 and 10. However, the bias caused by the genes with zero counts increases when value of ρ declines because the zero counts also induce bias in estimation of

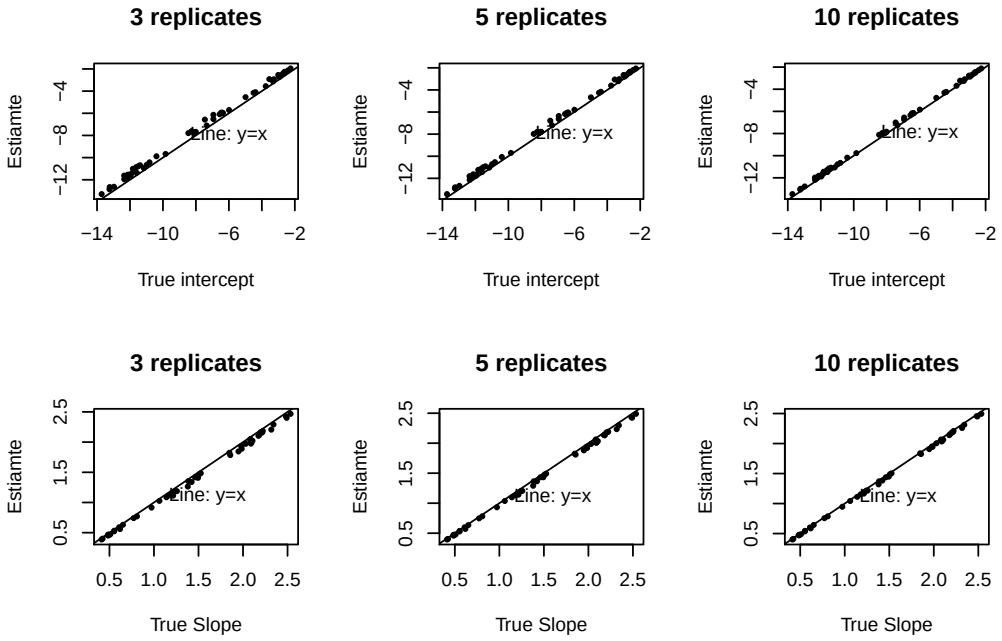

Figure 8: Simulation 2

Plotted true value versus average estimate of the ellipse's major axis
$y = PC_a + PC_b x$

platforms' covariance. Again by increasing the library size factor, the number of zero-count genes decreases and reduces the estimation bias.

Comparison with other estimator. The proposed estimator was also compared with others listed in Table 2 and the result is shown in Figure 9.

4.8.3 Simulation 3

Random model simulation without preprocessing error. In this simulation study, given a set of a and b, we first generated gene length and λ_i with the same method of Simulation 1 as the mean of gene i's distribution on both platforms. The latent RNA-Seq true expression measures for gene i and replicate j was $\lambda_{ij} \sim \Gamma(\lambda_i, \frac{\alpha_i}{\lambda_i})$ and $\frac{1}{\alpha_i} \sim \Gamma(0.85, 2)$ followed (Hardcastle and Kelly, 2010). Hence dispersion parameter caused by biological variation in gene i in RNA-Seq is $\frac{1}{\alpha_i}$. And $\mu_{ij} = \frac{log(lambda_{ij}) - a}{b}$. The range of those parameters are set similar to Simulation 1 and the generated genes were filtered according to Section 4.6. A total of 50 sets of parameters were generated. With each set of parameters, the number of replicates J was chosen at 3 levels: $(3, 5, 10)$ and 10 runs were repeated for each level for J. In total 1500 runs of simulation were generated. The estimates are shown in Table 4 and Figure 10.

Parameters					Estimate of intercept		Estimate of slope	
$E(\sigma_i^2)$	a	b	J	Lib factor	Mean	SD	Mean	SD
0.539	-9.16	1.789	3	4.243	-8.79	0.045	1.743	0.007
0.539	-9.16	1.789	5	4.243	-8.72	0.031	1.735	0.005
0.539	-9.16	1.789	10	4.243	-8.65	0.032	1.727	0.005
0.319	-4.56	0.778	3	2.717	-4.25	0.016	0.753	0.001
0.319	-4.56	0.778	5	2.717	-4.19	0.013	0.749	0.001
0.319	-4.56	0.778	10	2.717	-4.12	0.012	0.743	0.001
0.388	-5.32	0.941	3	3.222	-5.01	0.011	0.913	0.001
0.388	-5.32	0.941	5	3.222	-4.96	0.016	0.909	0.002
0.388	-5.32	0.941	10	3.222	-4.9	0.009	0.903	0.001
0.691	-7.67	1.48	3	1.263	-6.92	0.035	1.393	0.005
0.691	-7.67	1.48	5	1.263	-6.8	0.03	1.381	0.004
0.691	-7.67	1.48	10	1.263	-6.64	0.028	1.364	0.003
0.464	-6.48	1.17	3	3.232	-6.14	0.018	1.136	0.002
0.464	-6.48	1.17	5	3.232	-6.09	0.013	1.131	0.001
0.464	-6.48	1.17	10	3.232	-6	0.013	1.123	0.001

Table 4: Result of Simulation 3(truncated)

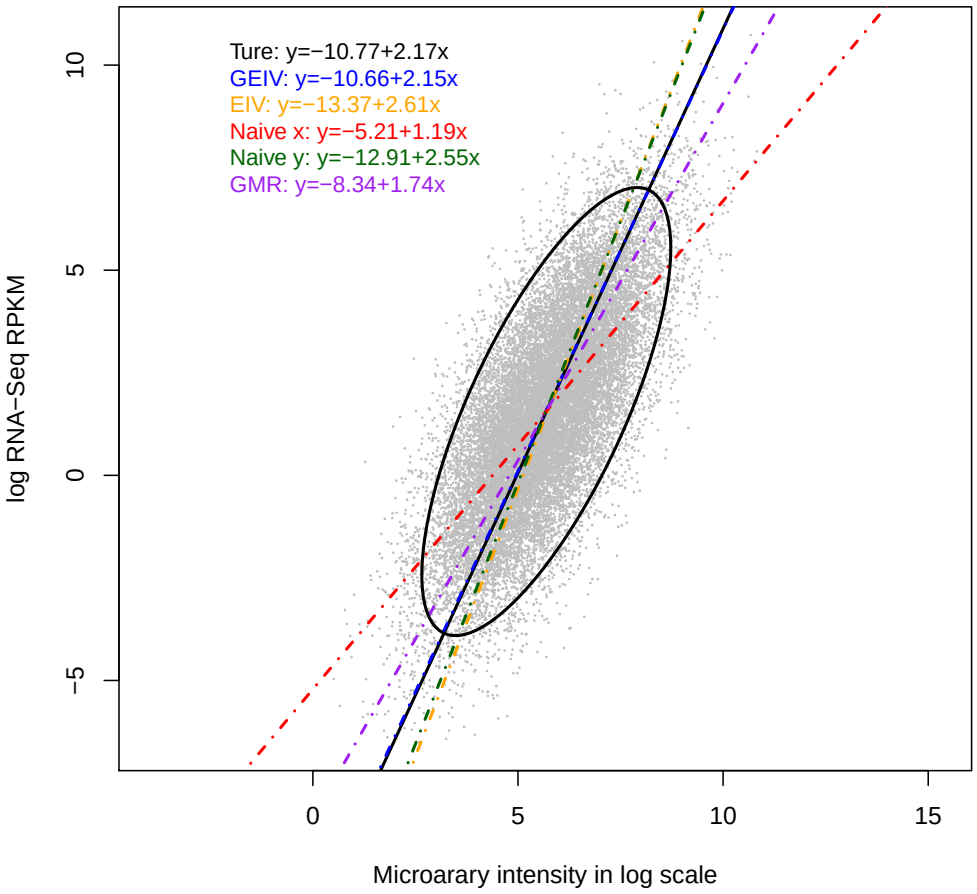

Figure 9: Comparison of estimators Simulation 2

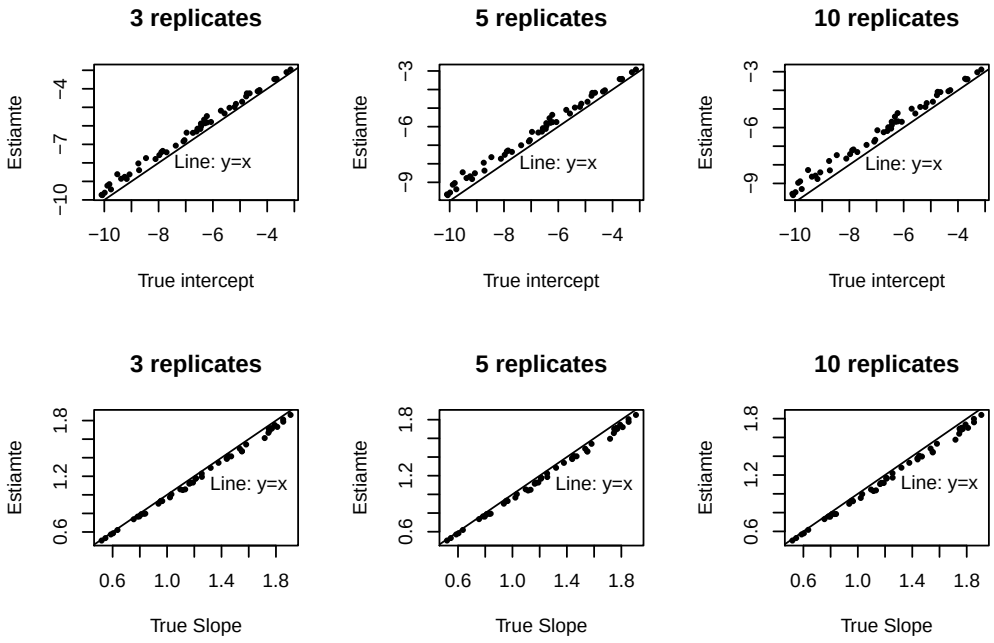

Figure 10: Simulation 3

Plotted true value versus average estimate

Simulation result. Same to shown Simulation 1, enlarging the sample size of replicates displays a minor reduction on bias and standard deviation of estimates, estimator in equation (22) shows convergence with each level of number of replicates, although is biased due to the number of zero-count genes. Again, increasing library size can reduce this bias.

Comparison with other estimator. The proposed estimator was also compared with others listed in Table 5 and the result in shown in Figure 11.

Estimator	Description
Naïve x	The GLM fit considering Microarray is exactly measured
Naïve y	The GLM fit considering RNA-Seq is exactly measured
GEIV	Proposed method
GMR	Geometric Mean of two Naïve estimators

Table 5: List of compared estimators in Simulation 3 and 4

4.8.4 Simulation 4

Random model simulation with preprocessing error. Given a set of a and b, we first generated gene length and μ_i and $log(\lambda_i)$ with the same method of Simulation 2 as $\xi_{i,array}$ and $\xi_{i,seq}$. Then we generated the biological replicates as $\lambda_{ij} \sim \Gamma(\alpha_i, \frac{\alpha_i}{\xi_{i,seq}})$ and $\mu_{ij} = \xi_{i,array} + \frac{log(\lambda_{ij}) - log(\xi_{i,seq})}{b}$. Similar to Simulation 2, the desired line here is no longer $y = a + bx$, but the major axis of the ellipse generated by above normal distribution: $y = PC_a + PC_b x$. ρ was randomly generated from $U[0.65 - 0.9]$ as in Simulation 2. Other parameters were the same with Simulation 1. Estimates of a and b were compared at different combination of a total of 50 sets of parameters were generated. With each set of parameters, the number of replicates J was chosen at 3 levels: $(3, 5, 10)$ and 10 runs were repeated for each level for J. In total 1500 runs of simulation were generated. The generated genes were filtered according to Section 4.6 and the estimate are shown in Table 6 and Figure 12.

Simulation result. Similar to shown simulations above, estimator shows convergence with a moderate number of replicates. Similar to Simulation

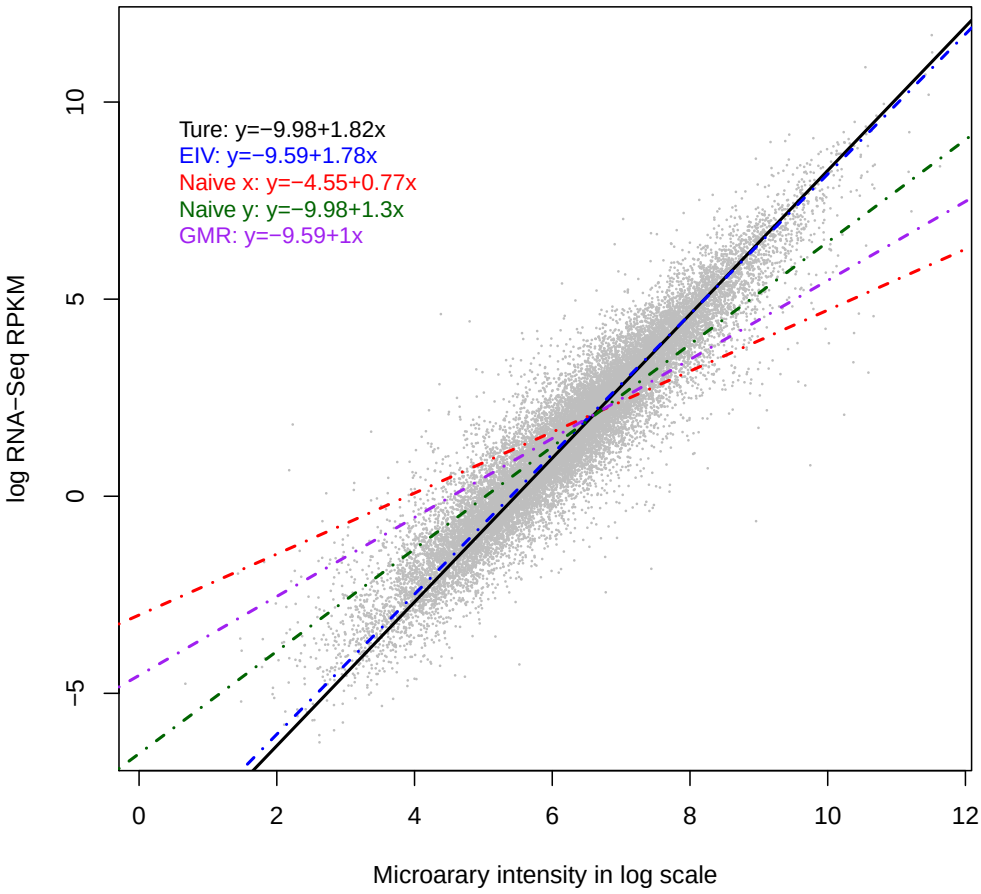

Figure 11: Comparison of estimators Simulation 3

Parameters								Esimates of a and b				Estimates of ellipse major axis			
$E(\sigma_i^2)$	a	b	J	E(lib)	ρ	PC_a	PC_b	$E(a)$	$sd(a)$	$E(b)$	sd(b)	$E(PC_a)$	$sd(PC_a)$	$E(PC_b)$	$sd(PC_b)$
0.46	-7.04	1.29	3	3.99	0.83	-7.49	1.36	-6.709	0.023	1.259	0.003	-7.115	0.029	1.32	0.004
0.46	-7.04	1.29	5	3.99	0.83	-7.49	1.36	-6.738	0.053	1.264	0.008	-7.149	0.064	1.326	0.009
0.46	-7.04	1.29	10	3.99	0.83	-7.49	1.36	-6.789	0.033	1.271	0.005	-7.212	0.037	1.335	0.006
0.72	-6.81	1.2	3	6.76	0.9	-6.98	1.23	-6.617	0.046	1.19	0.006	-6.778	0.051	1.214	0.007
0.72	-6.81	1.2	5	6.76	0.9	-6.98	1.23	-6.606	0.036	1.19	0.005	-6.769	0.04	1.214	0.005
0.72	-6.81	1.2	10	6.76	0.9	-6.98	1.23	-6.628	0.023	1.195	0.002	-6.796	0.027	1.219	0.003
0.53	-5.66	1.04	3	1.57	0.72	-5.77	1.06	-5.101	0.038	0.989	0.005	-5.069	0.051	0.984	0.007
0.53	-5.66	1.04	5	1.57	0.72	-5.77	1.06	-5.166	0.035	0.996	0.004	-5.153	0.047	0.994	0.006
0.53	-5.66	1.04	10	1.57	0.72	-5.77	1.06	-5.259	0.056	1.008	0.009	-5.281	0.081	1.011	0.012
0.65	-9.71	1.89	3	4.89	0.75	-11.93	2.26	-9.277	0.036	1.834	0.007	-11.4	0.077	2.191	0.013
0.65	-9.71	1.89	5	4.89	0.75	-11.93	2.26	-9.333	0.057	1.844	0.009	-11.51	0.102	2.211	0.017
0.65	-9.71	1.89	10	4.89	0.75	-11.93	2.26	-9.427	0.054	1.86	0.009	-11.62	0.064	2.23	0.011
0.44	-4.32	0.8	3	1.87	0.74	-3.91	0.75	-3.859	0.04	0.763	0.006	-3.335	0.049	0.692	0.007
0.44	-4.32	0.8	5	1.87	0.74	-3.91	0.75	-3.925	0.032	0.771	0.004	-3.425	0.036	0.703	0.004
0.44	-4.32	0.8	10	1.87	0.74	-3.91	0.75	-4.025	0.022	0.782	0.003	-3.552	0.03	0.716	0.005

Table 6: Result of Simulation 4 (truncated)

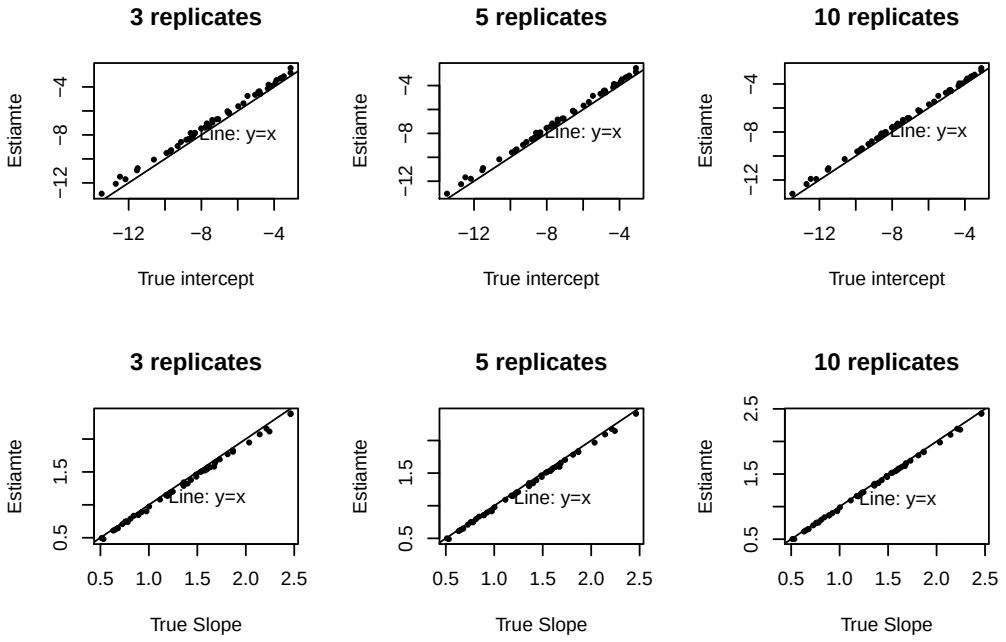

Figure 12: Simulation 4

Plotted true value versus average estimate of the ellipse's major axis
$y = PC_a + PC_b x$

2, the bias caused by the genes with zero counts increases when value of ρ declines because the biased estimated platforms' covariance also impacts on solution. Again by increasing the library size factor, the number of zero-count genes decreases and reduces the estimate bias.

Comparison with other estimator. The proposed estimator was also compared with others listed in Table 5 and the result in shown in Figure 13.

4.9 Real data analysis

Because in a real experiment, there is always a preprocessing error. In this section, we presented both the estimated two types of biases as well as the estimated ellipse and its major axis to describe the pattern in the expression profiles free of measurement error. Other estimators compared are listed in Table 7.

4.9.1 Human kidney

In this section, the proposed method was applied to parallel dataset with 3 technical replicates published in (Marioni et al., 2008). The procedure of generating Microarray and RNA-Seq are briefly introduced below according to the original paper.

Microarray description. Tissue samples from human kidney were collected and snap-frozen until processing, total RNA was extracted with TRIzol afterward. Aliquots from samples were hybridized to Affymetrix HG-U133 Plus 2.0 arrays with a single labeling reaction. After scanning, background-corrected, normalized intensities were obtained for all probe sets using the RMA algorithm (Gautier et al., 2004). Subsequently, mapping between probe sets and genes was conducted with information from NetAffx Analysis Center and BioMart (Flicek et al., 2008). Under multiple mapping, a single probe set were chosen randomly.

RNA-Seq description. Aliquots from the same RNA samples were sequenced with Illumina Genome Analyzer (32bp paired), following manufacturer protocol. Generated reads were aligned to Ensemble human genome 18 using ELAND. The library size is around 5 million reads per sample and

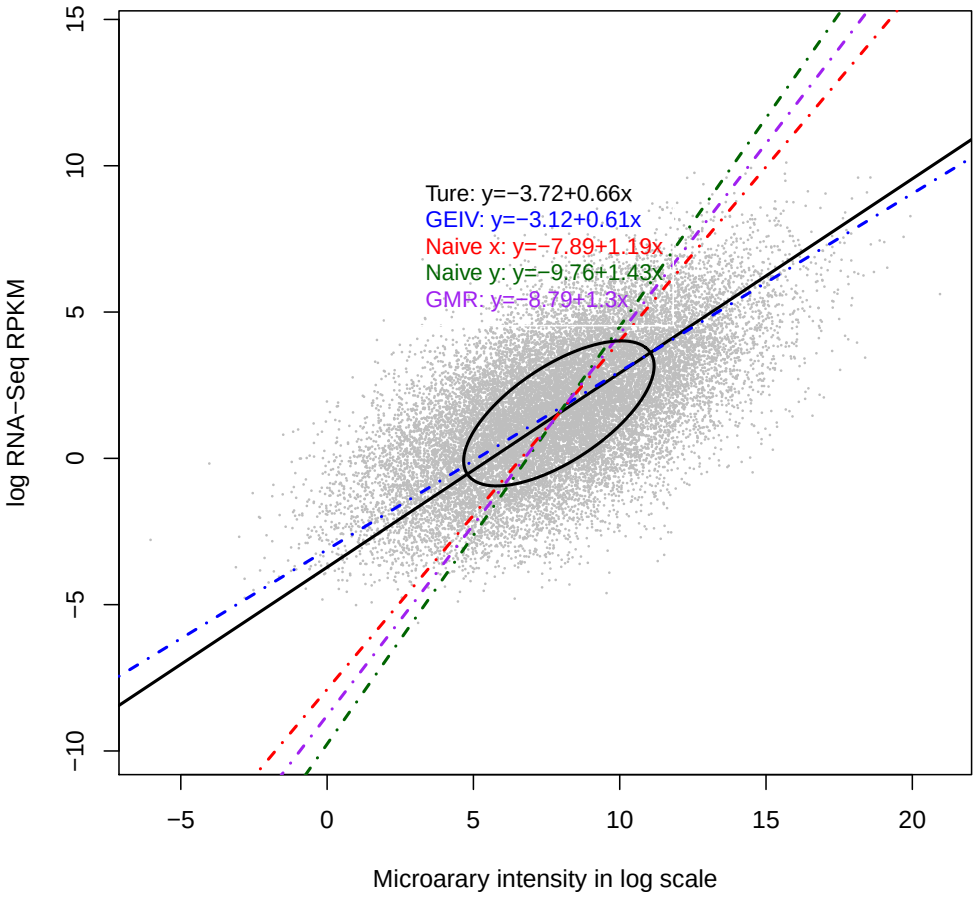

Figure 13: Comparison of estimators Simulation 4

40% of reads mapped uniquely to a genome location. The remaining reads were ignored.

Platform comparison. A reasonable assumption is there is no biological variation between technical replicates and the appropriate model is constant model in Section 4.7. The Pearson correlation between platforms is computed 0.64. Genes were filtered according to the method in Section 4.6 before comparison. From the result we can see the fitted ellipse agrees with the patten in genes with moderate and high expression, and shows some difference in genes with low RPKM, which is expected because the expectation of log RPKM value has a bias that is not negligible in low count genes as shown in Section 4.4. Specifically, $E(log(RPKM)) < log(E(RPKM))$, which agrees with the pattern shown in low count expressed genes. The estimators' comparison result is shown in Figure 14. The reference level is chosen as 90th percentile of each platform. The values in log scale are 6.33 for Microarray and 2.03 for RNA-Seq. Therefore, in log scale, derived 95% bootstrap confidence intervals (CI) of fixed bias is $[0.09, -0.18]$ and 95% CI of proportional bias is $[1.51, 1.55]$.

Estimator	Description
Naïve x	The GLM fit considering Microarray is exactly measured
Naïve y	The GLM fit considering RNA-Seq is exactly measured
GEIV error	Estimated $y = PC_a + PC_b x$
GEIV	Estimated $y = a + bx$
GMR	Geometric Mean of two Naïve estimators

Table 7: List of compared estimators in real data platform comparison

4.9.2 HT-29 cells

In this section, the proposed method is applied to parallel dataset published in (Xu et al., 2013) where the biological triplicates of HT-29 colon cancer cell lines are treated with: 1) dimethyl sulfoxide; 2) 5 μM 5-Aza as control and treatment groups. The procedure of generating Microarray and RNA-Seq are briefly introduced below according to the original paper and platform comparison is performed with both control and treatment groups.

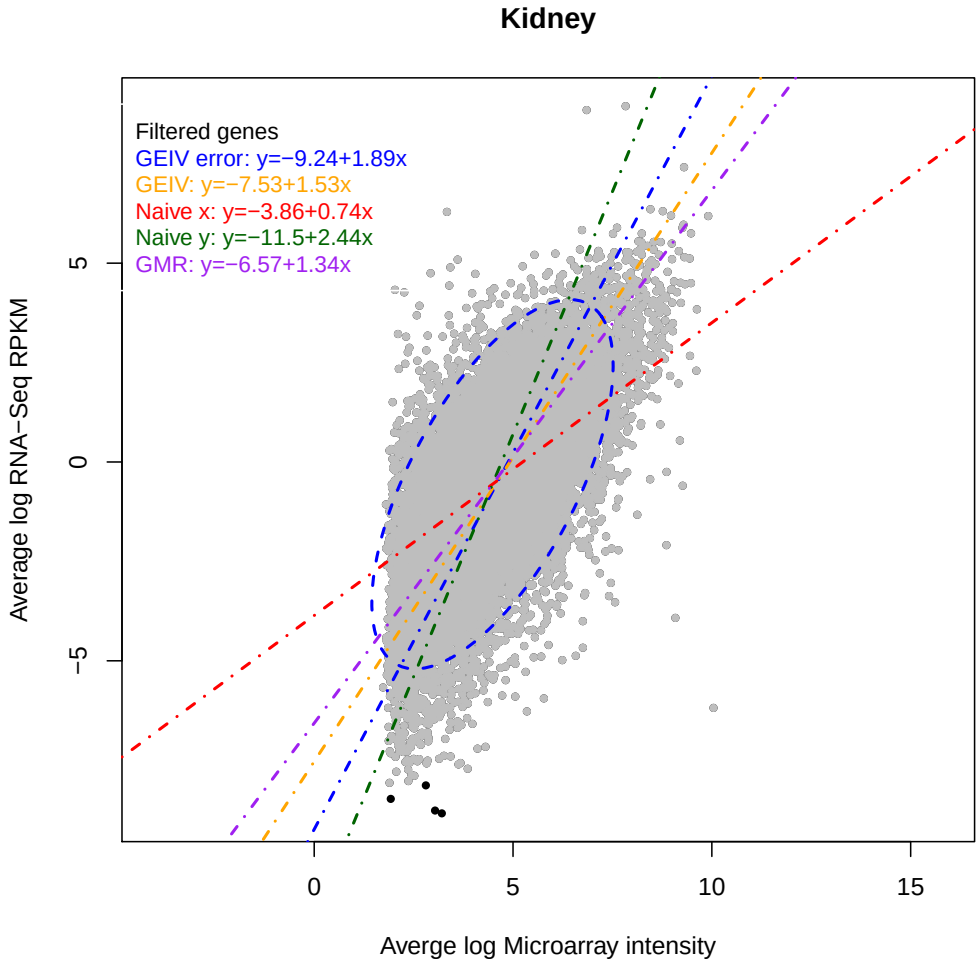

Figure 14: Comparison of platforms on human kidney data

Microarray description. Aliquots of RNA samples were hybridized to Affymetrix hgu133plus2.0 arrays, and then scanned in the according to the manufacturer's protocol. Background correction and normalization were done with Bioconductor's affy package with RMA algorithm and only retain the probes present in all three biological triplicates of either group. When multiple probes on the array map to a single gene, the probe with the max intensities was selected.

RNA-Seq description. Aliquots from the same RNA samples were subjected to Illumina 2000 sequencer (100bp paired). The RNA-Seq libraries were constructed using the TruSeq RNA Sample Preparation Kit (Illumina Inc., San Diego, CA). The library size is between 41 and 88 million reads per sample. Over 75 % of the short reads were aligned to Ensembl human genome 19 using Tophat (v2.0.1) (Trapnell et al., 2009) with default settings. HTSeq-count (Anders et al., 2014) were subsequently implemented to obtain the raw gene count of reads.

Platform comparison. A variation between biological replicates is expected and the platforms were compared with the random model in Section 4.7. The Pearson correlation between two platforms is reported around 0.68 (Xu et al., 2013). Genes were filtered according to the method in Section 4.6 before comparison. The observed patterns of fitted line and data points is similar to human kidney data. The estimators' comparisons for both groups are shown in Figure 15 and Figure 16 respective. With the same method with human kidney data to calculate expression reference in each platform, the references for this comparison are 7.13 for Microarray and 2.36 for RNA-Seq. 95% bootstrap CIs for fixed biases are $[0.23, 0.30]$ for control group and $[0.30, 0.35]$ for treatment group. 95% bootstrap CIs for fixed biases are $[1.10, 1.13]$ for control group and $1.14, 1.17]$ for treatment group.

4.9.3 Canine lymph nodes

In this section, the proposed method wass applied to parallel dataset published in (Mooney et al., 2013) where canine lymph nodes from 3 dogs without any treatment is measured with both Microarray and RNA-Seq as the control group. The procedure of generating Microarray and RNA-Seq are briefly introduced below according to the original paper except the RNA-Seq aligned

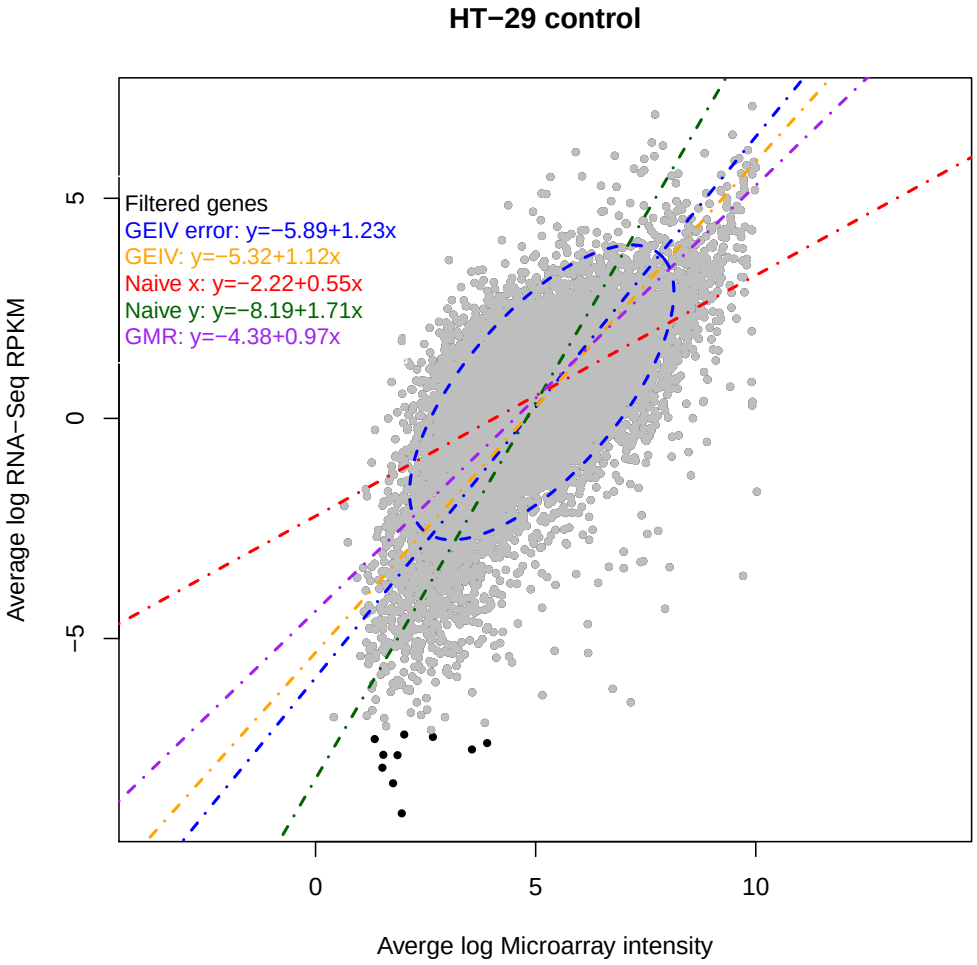

Figure 15: Comparison of platforms on HT-29 control group

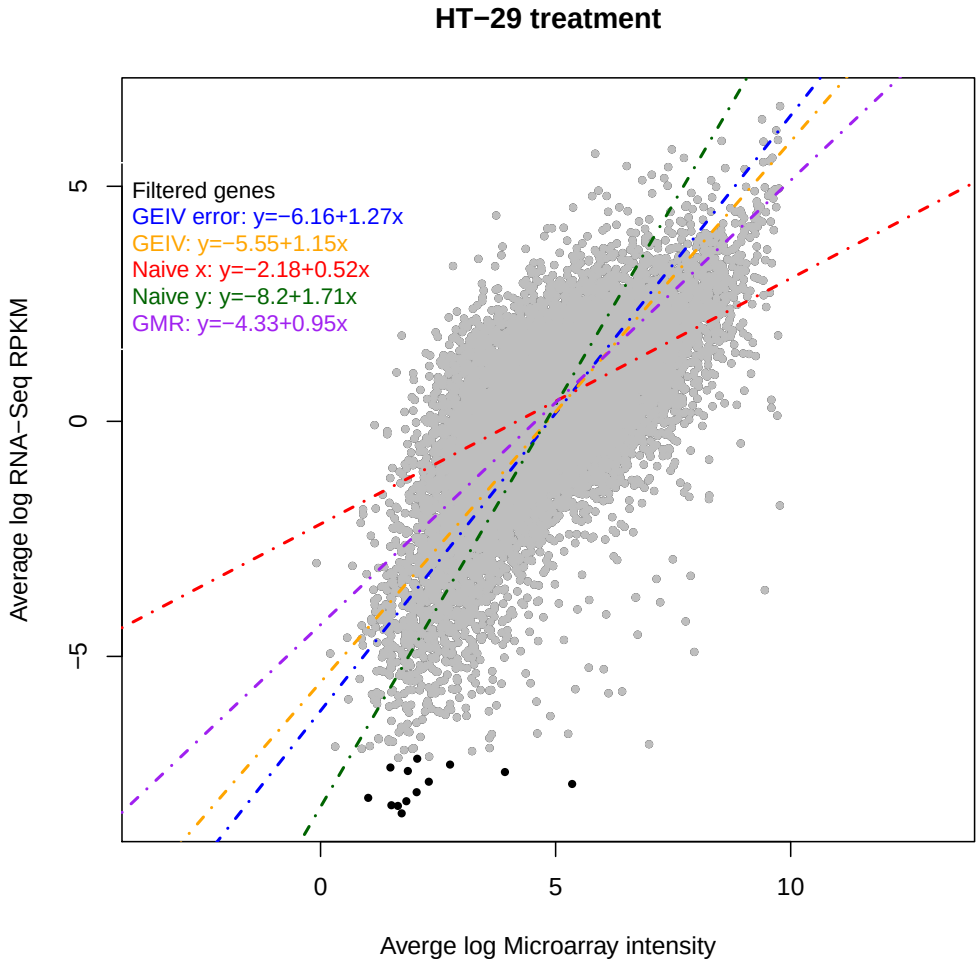

Figure 16: Comparison of platforms on HT-29 5 μM 5-Aza treatment group

fastq files are reprocessed with HT-Seq to obtain counts data for platform comparison.

Microarray description. RNA was preprocessed with the RNeasy Micro kit and $WT - Ovation^{TM}$ Pico RNA Amplification System (NuGen Technologies, Inc) and measured using GeneChip Canine Genome V2.0 Array (Affymetrix). Obtained intensity data was normalized using the MAS 5.0 algorithm within the Affymetrix Gene Console. Further quality check was done after analyzing distribution with R Affy package, afterwards mapping between probe sets and genes was conducted with BioMart.

RNA-Seq description. Same RNA is sequenced using the Illumina HiSeq 2000 and the generated library sizes are between 74 and 134 million reads per sample. In average 65% of the short reads were aligned to Ensemble dog genome canFam3 using Tophat (v2.0.1) with default settings. HTSeq-count (Anders et al., 2014) were subsequently implemented obtain the raw gene count of reads.

Platform comparison. A variation between different subjects is expected and the platforms were compared with the random model in Section 4.7. The Pearson correlation between two platforms is reported at 0.70. Genes are filtered according to the method in Section 4.6 before comparison. Patterns similar to above comparisons were observed. The estimators' comparison result are shown in Figure 17. With the same method to calculate expression reference in each platform, the references for this comparison are 7.89 for Microarray and 4.78 for RNA-Seq. 95% bootstrap CI of fixed bias is $[-0.28, -0.20]$ and 95% bootstrap CI of fixed bias of proportional bias is $[1.07, 1.12]$.

4.10 Discussion

The correlation based quantification of platform consistency is further developed into fitting a line between expression profiles from Microarray and RNA-Seq platforms. Differences between the two platforms show both fixed and proportional biases detected by the generalized linear errors-in-variable (GEIV) regression model, especially with more dated RNA-Seq technology

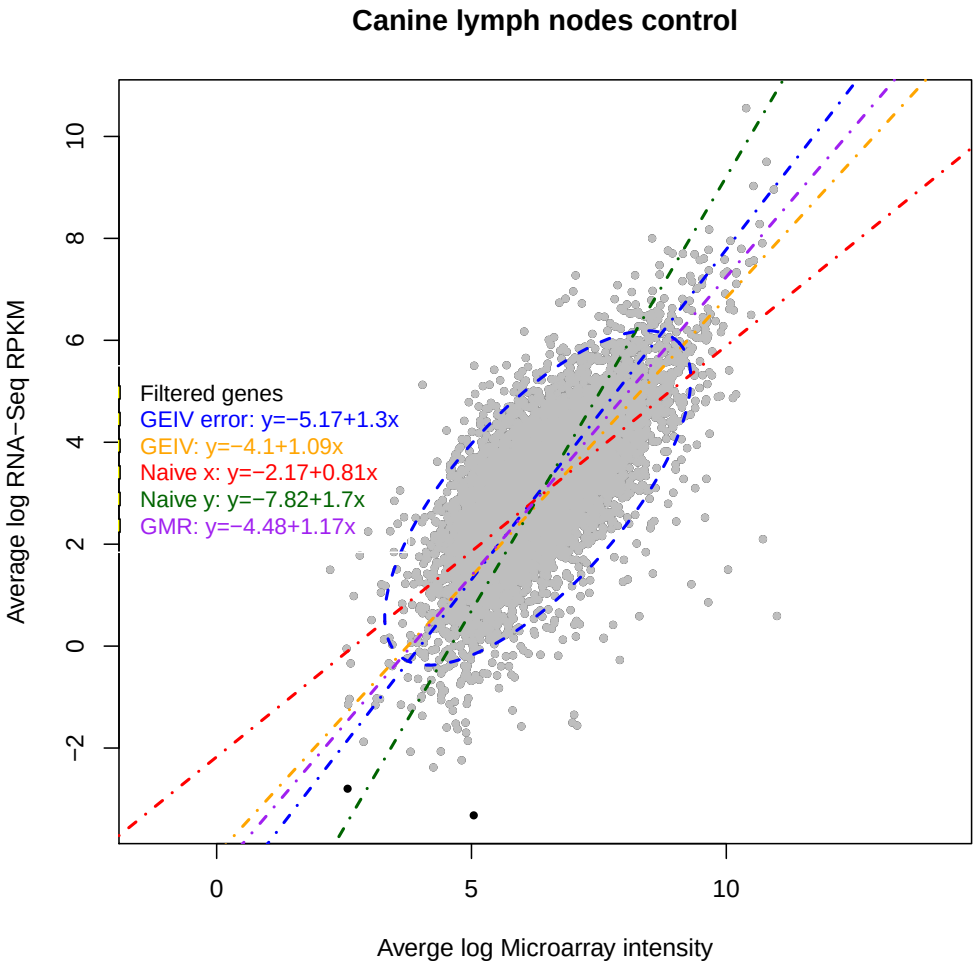

Figure 17: Comparison of platforms on Canine lymph nodes

and alignment algorithm used in Section 4.9.1. However, with more recent technology, we observe only a minor bias and the magnitude of bias remains consistent across three datasets shown in Section 4.9.2 and 4.9.3, which demonstrates even the Microarray is the platform closer to underlying truth, RNA-Seq is still an appealing measurement method since it requires no per-defined probe arrays and can detect novel genes and mutation. Consequently, the bias shown in this study is not necessarily a result of different quantification theory of gene expression and with technology and data summarization algorithms developing, it can be expected the bias will be further reduced (see Table 8).

Table 8: Summary of real data comparison

	Kidney	HT29 control	HT29 treatment	Canine
Aligned reads [million]	1.8-2.1	36-79	36-79	55-90
Gene number	15406	11960	11761	6548
Library type	36bp paried	100bp paired	100bp paired	100bp paired
Pearson correlation	0.64	0.68	0.68	0.7
95% CI for fixed bias	[0.09,0.18]	[0.23,0.30]	[0.3,0.35]	[-0.28,-0.20]
95% CI for proportioal bias	[1.51,1.55]	[1.10,1.13]	[1.14,1.17]	[1.07,1.12]

5 Comparison of identified changes in gene expression measured by Microarray and RNA-Seq

5.1 Background

This section includes the contribution of the dissertation's author from published work (Xu et al., 2013), which introduces comparing significant differentially expressed genes (DEG) detected by Microarray and RNA-Seq based on the HT29 dataset previously introduced in 4.9.2. Cultures of HT-29 colon cancer cell line was maintained and evenly divided into 2 groups consisted of three of 150 mm cultures were treated with: 1) dimethyl sulfoxide (vehicle alone, 0 μM 5-Aza) and 5 μM 5-Aza for five days. It has been reported 5-Aza alters mRNA gene expression in HT-29 cells (Cheetham et al., 2008). RNA samples were extracted separately to generate parallel RNA-Seq, Microarray and qRT-PCR data. Several designed tests on DEGs for both RNA-Seq and Microarray data, were compared with both synthetic and real dataset. Selected genes that not detected by one of the technologies were further validated using qRT-PCR assays, In additional,Ingenuity Pathway Analysis (IPA) was applied to assess overlap of significant pathways altered with the treatment of 5-aza-deoxy-cytidine. A major rationale for including this analysis is to further validate the consistency between Microarray and RNA-Seq by comparing the statistical significant changes between two biological conditions identified from both platforms.

5.2 DEG algorithms for Microarray and RNA-Seq data

DEG lists for Microarray were generated using 3 different algorithms: T-test with Benjamini-Hochberg correction (Benjamini and Hochberg, 1995), SAM (Tusher et al., 2001) and eBayes (Smyth et al., 2004) for two comparisons: 1) 5μM vs. 0μM 5-Aza groups and 2) 10μM vs. 5μM 5-Aza groups, respectively. And similarly the SAMSeq (Li and Tibshirani, 2013), DESeq (Anders and

Huber, 2010), baySeq (Hardcastle and Kelly, 2010) algorithms were applied to generate DEG list for RNA-Seq data. In addition, a cutoff (fold change $>$ 2) which is empirically the minimal change with biological significance was applied to DEG lists.

5.3 Simulation study

Different from simulation in 4.8 where corrected and normalized Microarray data is simulated, the simulation strategy was to generate synthetic raw data based on human kidney published in (Marioni et al., 2008). The simulated datasets for Microarray and RNA-Seq have same magnitude of changes between groups. Generated dataset was used to choose the algorithms with most power under the same cutoff: false discovery rates (FDR) < 0.05 and fold change > 2. The selected method's DEG list was considered the DEG detected by corresponding platform to adjust the effect of sensitivity and specificity differences in algorithms.

The Microarray data was simulated with the model described previously by (Rocke and Durbin, 2001), in which raw gene expression is modeled as $x = \alpha + e^{\mu+\delta} + \nu$, where α is the mean background bias, δ, ν are normal error term and μ is true intensity.

Since it is well known Microarray among subjects variance varies with level of expression. The range of parameters was determined through a variance stabilizing transformation (Durbin et al., 2002) with the method described in (Xu et al., 2013). The RNA-Seq raw counts were generated from similar synthetic data simulation method described in (Kvam et al., 2012) with the kidney dataset as reference data instead.

Preset DEGs were randomly selected and the log fold changes of these preset DEGs were generated from a normal mixture distribution consist of a component for up regulated genes and down regulated genes with the same parameter setting in (Kvam et al., 2012). The sensitivity and FDR were evaluated for under different mean magnitude of simulated fold changes and cutoffs. The result shows SAM and DESeq shows the overall largest AUC of ROC curve and hence they were selected to represent each platform.

5.4 Cross-platform comparison of DEG lists

The results in the real dataset are, treatment group of HT-29 cells with 5 μM 5-Aza shows alternation of gene expression levels by both up-regulation

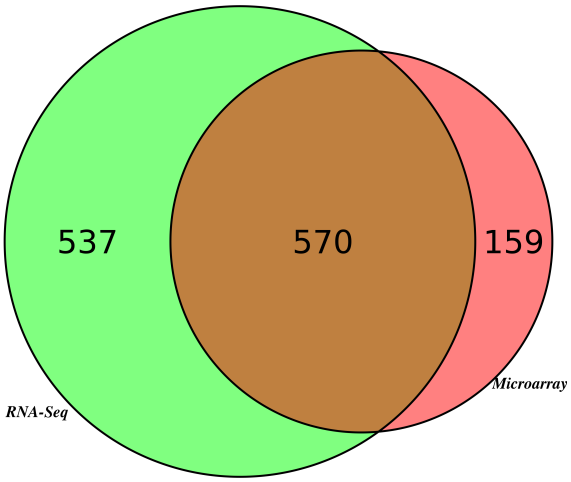

Figure 18: Identified commonly detectable DEG from Microarray and RNA-Seq

(↑) and down-regulation (↓) of genes. SAM identified 794 ↑ and 256 ↓ and DESeq found 1840 ↑ and 300 ↓ (Xu et al., 2013). Additionally the increasing of SPARC gene expression, which is previously reported (Cheetham et al., 2008) after similar treatment, is only identified by RNA-Seq. The result is consistent with biological effect of 5-Aza treatment. Note that only 11960 genes as 'commonly detectable' (Xu et al., 2013), where other genes are 1) not on Microarray chip 2) measured with extremely low expression profile in either platforms. And the number of those genes identified by either platform is shown in Figure 18. It is found that 78.2% of Microarray identified commonly detectable DEG is overlapped with RNA-Seq. Among the total 1266 genes shown in Figure 18, only 12.6% is not detected in RNA-Seq. It is shown in Figure 19, the commonly detectable DEGs only identified by RNA-Seq shows less overall measured fold change comparing those show significance in both platforms.

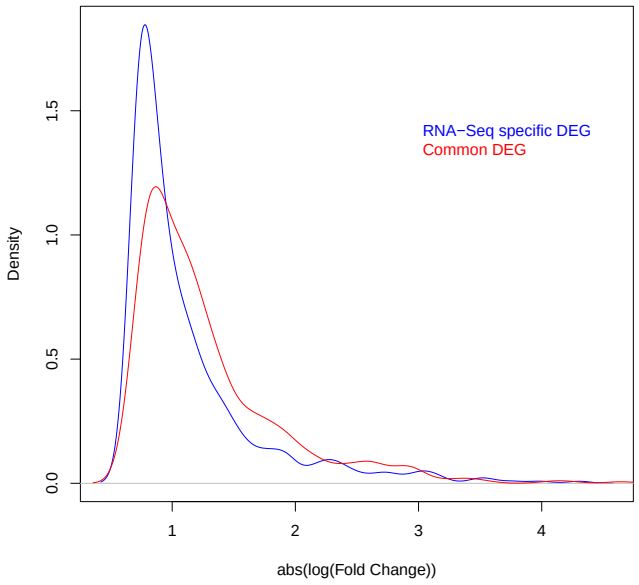

Figure 19: Fold change distribution of DEG identified by RNA-Seq

5.5 Comparison in DEG functions and qRT-PCR confirmation

The functional analysis of DEG is performed with ingenuity pathway analysis (IPA®, QIAGEN Redwood City,www.qiagen.com/ingenuity).33 IPA canonical pathways in both Microarray and RNA-Seq, while 152 pathways were only able to be identified in RNA-Seq and no pathways is only significant on Microarray (Xu et al., 2013). In additional, the 'gold standard' technology in gene expression, qRT-PCR was applied to confirm selected DEGs. Overall qRT-PCR validation is shown in Table 9.

	Both	RNA-Seq only	Microarray only
Confirmed	3	1	3
Not confirmed	1	3	2

Table 9: PCR validation of DEG

5.6 Discussion

The comparison on DEG from both platforms shows RNA-Seq agrees with the majority of significant changes on Microarray. Since the RNA-Seq specific DEGs are confirmed by qRT-PCR, the result also shows RNA-Seq is more sensitive to smaller changes which makes it able to identify more DEGs and biological pathways. It is consistency with the GEIV model fitted solely on measurements in Section 4.9 as estimated slope is greater than 1.

6 Pairwise EIV and SEM linear model with 16S sequencing on different regions of 16S ribosomal RNA

Even the distributions is assumed to be normal, the challenges to apply EIV model to comparison of two platforms is acquiring enough replications on same subjects. However, when they are available, we can apply the method proposed in (Barnett, 1970) to fit a linear relationship between them. On the other hand, SEM is relative easier to implement when there are three or mode sources of measurements to compare. The section shows a study of consistency of pairwise EIV and SEM based on 16S sequencing on three different regions of 16s ribosomal RNA.

6.1 Background

Genes on nine hypervariable regions (V1-V9) on Bacterial 16S ribosomal RNA have a considerable ability to differentiate bacteria (Chakravorty et al., 2007). Although classification based on 16S ribosomal RNA does not completely collapse into the biological classification of bacteria, the obtained operational taxonomic unit (OTU) has been demonstrated to be the practical grouping of bacteria in different natural levels like phylum, order and genus.Therefore, an approach to quantify the microbial abundance is to quantify the expression of those genes with sequencing. However, usually only a selected part of those nine regions are measured and 3 common choices are V1V2,V1V3 and V3V4. In this study, microbial abundance measured from those 3 regions are compared with a SEM model following (Wu et al., 2013) and a pairwise EIV model.

6.2 Data structural and model

Six different mouses was measured with 16S sequencing with V1V2,V1V3 and V3V4 primers. Each sample was amplified 10 times with different barcoded primers, which were essentially 10 replicates. For a specific mouse, the

obtained data is in a form of count data, and the structure is shown in Table 10.

	V1V2	V1V3	V3V4
OTUs	repliates 1-10	repliates 1-10	repliates 1-10
Bacteria/Actinobacteria	counts	counts	counts
Bacteria/Bacteroidetes	counts	counts	counts
Bacteria/Proteobacteria	counts	counts	counts
...	...	...	...

Table 10: Data structural of abundance data, phylum-subphylum level

6.3 Method

The count data were converted to percentage and then subject to a arcsin square root transformation to generate normalized abundance of each OTU. For a specific OTU, since the transformation applied here is to approximate proportional data to a normal distribution (McDonald, 2009), the obtained abundance was considered to be normal variable. And a SEM model can be written as:

$$\begin{aligned} x_{ij} &= \xi_i + \delta_{ij} \\ y_{ij} &= a_y + b_y \xi_i + \epsilon_{ij} \\ z_{ij} &= a_z + b_z \xi_i + \omega_{ij} \,. \end{aligned} \tag{27}$$

The notations are: $i = 1, 2..., I$ stands the index for subject and $j = 1, 2, ..., J$ stands for replicates, x, y and z stands for V1V2, V1V3 and V3V4 measurements respectively while ξ_i is the latent true abundance. We assume $\xi_{ij} \sim N(\mu, \tau^2)$, $\delta_{ij} \sim N(0, \sigma_\delta^2)$, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, $\omega_{ij} \sim N(0, \sigma_\omega^2)$ and the random variables are independent.

It is trivial to show the sample means is sufficient statistics for the underlying true abundance in this model. With notations: $\mathbf{T} = \begin{pmatrix} \bar{x}_{i.} \\ \bar{y}_{i.} \\ \bar{z}_{i.} \end{pmatrix}$, $\mathbf{b}' = \begin{pmatrix} 1 \\ b_y \\ b_z \end{pmatrix}$ and $\mathbf{r} = \begin{pmatrix} \bar{\delta}_{i.} \\ \bar{\epsilon}_{i.} \\ \bar{\omega}_{i.} \end{pmatrix}$. It can be shown covariance matrix of replicates' means is

$\mathbf{\Lambda} = E(\mathbf{TT}') = \mathbf{bb}'\tau^2 + \mathbf{R}$ where $\mathbf{R} = \frac{1}{j} diag(\sigma_\delta^2, \sigma_\epsilon^2, \sigma_\omega^2)$. Hence the SEM model can be solved with same method in Section 2.3 after averaging the replicates.

A linear EIV comparison can be also constructed using any two equations in model (27). Because the solution of a structural EIV suffers from multiple root and stability issues, we assume a functional model shown and the variance measurement error for different subjects at a same region is the same. The solution of this model is discussed in (Barnett, 1970) as a special case of the solution shown in Section 2.2.

$$
\begin{aligned}
x_{ij} &= \xi_i + \delta_{ij} \\
y_{ij} &= \eta_i + \epsilon_{ij} \\
\eta_i &= a + b\xi_i \\
\delta_{ij} &\overset{i.i.d}{\sim} N(0, \sigma_\delta^2) \\
\epsilon_{ij} &\overset{i.i.d}{\sim} N(0, \sigma_\epsilon^2)\,.
\end{aligned}
$$

Since 3 of the mouse has APC genotype and 3 wild genotype, to employ maximum samples, a repeated measures ANOVA was first fitted to access the significance of each OTU's difference between APC and wild group. Lachnospiraceae was selected as the OTU to compare since it has moderate abundance and large p-value (> 0.6). V1V3 was chosen to be the reference (slope=1 in SEM and EIV) because it overlaps with both two other regions. 4 libraries with total counts $< 10{,}000$ were excluded based on quality control suggested by the lab performing the experiment. An subset of replicates for each sample was selected to balance the experimental design (see Appendix 7.3).

6.4 Result

SEM on Lachnospiraceae. The result is shown in Table 11. We can observe minor proportional bias and fixed bias comparing both V1V2 and V3V4 to V1V3. In addition, V3V4 is more consistent with V1V3 comparing to V1V2.

EIV on Lachnospiraceae. The result is shown in Table 12. Same to SEM, We can observe minor proportional bias and fixed bias comparing both V1V2

	V1V2	V1V3	V3V4
slope	1.3	1	1.09
intercept	-0.19	0	-0.15
reliability	0.98	0.92	0.93

Table 11: SEM estimates on Lachnospiraceae

and V3V4 to V1V3. Furthermore, V3V4 is still more consistent with V1V3 comparing to V1V2. The fitted line with V1V3 as x is shown at Figure 20.

	V1V3(x) versus V1V2(y)	V1V3(x) versus V3V4(y)	V1V2(x) versus V3V4(y)
slope	1.28	1.14	0.87
intercept	-0.28	-0.05	0.22

Table 12: EIV estimates on Lachnospiraceae

Discussion. Since both SEM and EIV model are based on linear model, we expected the pairwise EIV will agree with the estimates from corresponding regions in SEM and from the result we can see, with the same data, EIV and SEM only show a trivial difference. It supports to use EIV as an alternative approach when only comparing two platforms where SEM is no applicable.

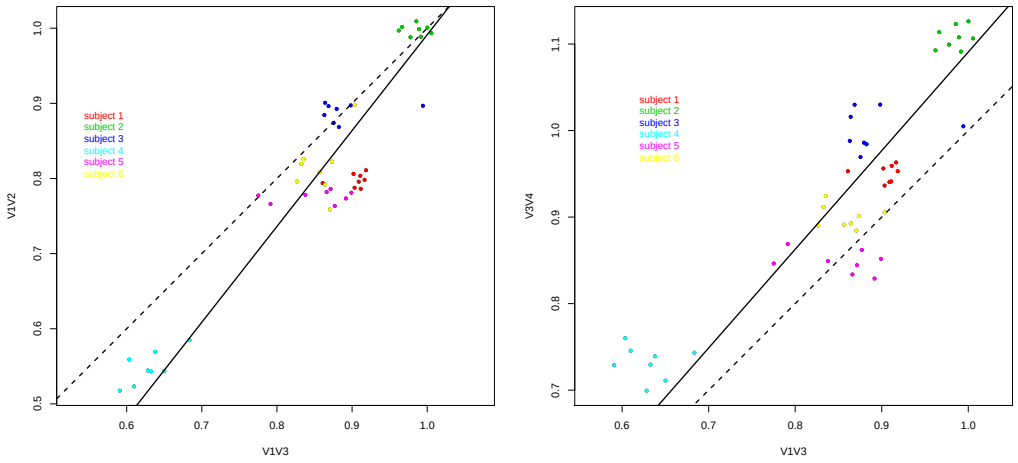

Figure 20: Fitted EIV line: V1V2 and V3V4 compare to V1V3

The line is fitted EIV line and the dotted line is reference y=x.

7 Appendix

7.1 Proof

Proof of Lemma 4.1.

$$g(\bar{x}_{i.};\theta) = \begin{bmatrix} \bar{x}_{i.} - \frac{log(\lambda_i)-a}{b} \\ \bar{x}_{i.}^2 - \bar{\delta}_{i.}\bar{x}_{i.} - (\frac{log(\lambda_i)-a}{b})^2 \end{bmatrix}.$$

It follows,

$$E(g(\bar{x}_{i.};\theta)) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and

$$V_i = Var(g(\bar{x}_{i.};\theta)) = \begin{bmatrix} \frac{\sigma_i^2}{J} & \frac{\mu_i\sigma_i^2}{J} \\ \frac{\mu_i\sigma_i^2}{J} & \frac{\mu_i^2\sigma_i^2}{J} \end{bmatrix}.$$

Empirically we can assume μ_i and σ_i satisfying $tr([V'^{1/2}V_{ij}^{1/2}) < A$ for all i,j and some A. Then when I goes to infinity, denote $\lim \frac{\sum_i V_i}{I} = V$. by special

case of Lindeberg-Feller central limit theorem stated in (Demidenko, 2013), $\sqrt{I}f(\theta) \xrightarrow{d} N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, V)$.

Proof of Lemma 4.2. From delta method, let $y_{i.} = \Sigma_j y_{ij}$ and $N_{i.} = \Sigma_j N_{ij}$. We have

$$logy_{i.} \ - logN_{i.} \ = log(\lambda_i) - \frac{1}{2\lambda_i N_{i.}} + o(\frac{1}{\lambda_i N_{i.}})$$

$$E\left(\frac{1}{y_{i.}+1}\right) = \frac{1}{\lambda_i N_{i.}}(1 - e^{-N_{i.}\lambda_i}) \, .$$

From LLN,

$$\Sigma_i \left(log\frac{y_{i.}}{N_{i.}} \ + \frac{1}{2\,(y_{i.}+1)} \right) \to \Sigma_i E \left(log\frac{y_{i.}}{N_{i.}} \ + \frac{1}{2\,(y_{i.}+1)} \right) \approx \Sigma_i log\lambda_i \ .$$

From delta method,

$$var\left(log\frac{y_{i.}}{N_{i.}} \ + \frac{1}{2\,(y_{i.}+1)} \right) \approx \frac{1}{N_{i.}\lambda_i}$$

$$E\left(log\frac{y_{i.}}{N_{i.}} \ + \frac{1}{2\,(y_{i.}+1)} \right)^2 - var\left(log\frac{y_{i.}}{N_{i.}} \ + \frac{1}{2\,(y_{i.}+1)} \right)$$

$$= \left[E\left(log\frac{y_{i.}}{N_{i.}} \ + \frac{1}{2\,(y_{i.}+1)} \right) \right]^2 \approx \Sigma_i (log\lambda_i \ )^2 \, .$$

From LLN,

$$\Sigma_i \left[\left(log\frac{y_{i.}}{N_{i.}} \ + \frac{1}{2\,(y_{i.}+1)} \right)^2 - \frac{1}{(y_{i.}+1)} \right] \approx \Sigma_i (log\lambda_i \ )^2$$

$$\Sigma_i \frac{(x_{ij} - \overline{x_{i.}})^2}{J-1} \to \Sigma_i \sigma_i^2 \, .$$

Proof of Lemma 4.3. Assuming $p(x_{ij}, y_{ij})$ is the unknown density function for (x_{ij}, y_{ij}). We have

$$\iint E(\mu_{ij}(\lambda_{ij})|x_{ij}, y_{ij}) p(x_{ij}, y_{ij}) dx_{ij} dy_{ij} = E(\mu_{ij}(\lambda_{ij})).$$

Denote $Z_{ij} = E(\mu_{ij}(\lambda_{ij})|x_{ij}, y_{ij})$. Then we have

$$E(Z_{ij}) = E_{\lambda_{ij}}(\mu_{ij}(\lambda_{ij}))$$
$$Var(Z_{ij}) \leq Var_{\lambda_{ij}}(\mu_{ij}(\lambda_{ij})) = \frac{\psi_1(\alpha_i)}{b^2}\,.$$

Where ψ is the digamma function and ψ_1 is trigamma function. Therefore, as $I \to \infty$, if α_i and σ_i satisfy $Var(\frac{\Sigma_i\Sigma_j Z_{ij}}{IJ\sigma_i^2}) \to 0$,

$$\frac{\Sigma_i\Sigma_j Z_{ij}}{IJ\sigma_i^2} \to \frac{\Sigma_i\Sigma_j E_{\lambda_{ij}}(\mu_{ij}(\lambda_{ij}))}{IJ\sigma_i^2} = \frac{\Sigma_i\Sigma_j(\psi(\alpha_i) - log\beta_i - a)}{IJb\sigma_i^2}\,.$$

Similarly, if we denote $Z'_{ij} = E_{\lambda_{ij}}(\mu_{ij}^2(\lambda_{ij})|x_{ij}, y_{ij})$. We want to show that under regularity conditions $Var(\frac{\Sigma_i\Sigma_j Z'_{ij}}{IJ\sigma_i^2}) \to 0$.

$$\Sigma_i\Sigma_j Var(Z'_{ij}) = \Sigma_i\Sigma_j[E(Var(Z'_{ij}|x_{ij})) + Var(E(Z'_{ij}|x_{ij}))]\,.$$

By law of total conditional variance, we have:

$$Var(Z'_{ij}|x_{ij})) \leq Var(\mu_{ij}^2(\lambda_{ij})|x_{ij}) = 4x_{ij}^2\sigma_i^2 + 2\sigma_i^4$$

$$E(Z'_{ij}|x_{ij}) = E(E_{\lambda_{ij}}(\mu_{ij}^2(\lambda_{ij})|x_{ij}, y_{ij})|x_{ij}) = E(\mu_{ij}^2(\lambda_{ij})|x_{ij}) = x_{ij}^2 + \sigma_i^2$$

$$\Sigma_i\Sigma_j Var(Z'_{ij}) \leq \Sigma_i\Sigma_j[E(4x_{ij}^2\sigma_i^2 + 2\sigma_i^4) + Var(x_{ij}^2)]\,.$$

It follows if $\Sigma_i\Sigma_j\frac{[E(4x_{ij}^2\sigma_i^2+2\sigma_i^4)+Var(x_{ij}^2)]}{(IJ\sigma_i^2)^2} \to 0$,

$$\Sigma_i\Sigma_j E_{\lambda_{ij}}(\frac{\mu_{ij}^2(\lambda_{ij})}{IJ\sigma_i^2}|x_{ij}, y_{ij}) \xrightarrow{p} \Sigma_i\Sigma_j E_{\lambda_{ij}}(\frac{\mu_{ij}^2(\lambda_{ij})}{IJ\sigma_i^2})$$
$$= \Sigma_i\Sigma_j(\frac{\psi_1(\alpha_i) + [(\psi(\alpha_i) - log\beta_i) - a]^2}{IJb^2\sigma_i^2})\,.$$

Proof of Lemma 4.4. Similar to **Lemma 4.2**. If we denote $\delta_{ij} = x_{ij} - \mu_{ij}$, we have

$$\begin{aligned}
&\Sigma_i\Sigma_j E_{\lambda_{ij}}(\frac{x_{ij}\mu_{ij}(\lambda_{ij})}{IJ\sigma_i^2}|x_{ij}, y_{ij}) = \Sigma_i\Sigma_j \frac{x_{ij}^2}{IJ\sigma_i^2} - \Sigma_i\Sigma_j \frac{E(x_{ij}\delta_{ij}|x_{ij}, y_{ij})}{IJb\sigma_i^2} \\
&\xrightarrow{p} \Sigma_i\Sigma_j \frac{x_{ij}^2}{IJ\sigma_i^2} - \Sigma_i\Sigma_j \frac{E(x_{ij}\delta_{ij})}{IJ\sigma_i^2} = \Sigma_i\Sigma_j \frac{x_{ij}^2}{IJ\sigma_i^2} - 1\,.
\end{aligned}$$

Note that, this is an alternative approximation of $\Sigma_i\Sigma_j E_{\lambda_{ij}}(\frac{x_{ij}\mu_{ij}(\lambda_{ij})}{IJ\sigma_i^2}|x_{ij}, y_{ij})$, because it converges $\Sigma_i\Sigma_j E_{\lambda_{ij}}(\frac{\mu_{ij}^2(\lambda_{ij})}{IJ\sigma_i^2})$ as well. In this way, we can construct a valid estimate equation with two terms that are expected to be the same given right a and b.

Proof of Lemma 4.5. We define.

$$g(x_{ij};\theta) = \begin{bmatrix} x_{ij} - \frac{\psi(\alpha_i) - log\beta_i - a}{b} \\ x_{ij}^2 - \delta_{ij}x_{ij} - (\frac{\psi_1(\alpha_i) + [\psi(\alpha_i) - log\beta_i - a]^2}{b^2}) \end{bmatrix}.$$

It follows,

$$E(g(x_{ij};\theta)) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and

$$\begin{aligned}
&V_{ij} = Var(g(x_{ij};\theta)) \\
&= \begin{bmatrix} Var(\mu_{ij}) + \sigma_i^2 & Cov(\mu_{ij}^2, \mu_{ij}) + \sigma_i^2 E(\mu_{ij}) \\ Cov(\mu_{ij}^2, \mu_{ij}) + \sigma_i^2 E(\mu_{ij}) & Var(\mu_{ij}^2) + E(\mu_{ij}^2)\sigma_i^2 \end{bmatrix}.
\end{aligned}$$

From

$$
\begin{aligned}
log\lambda_{ij} &= a + b\mu_{ij} \\
Var(log\lambda_{ij}) &= \psi_1(\alpha_i) + \sigma_i^2 \\
Cov[(log\lambda_{ij})^2, log\lambda_{ij}] &= \frac{\Gamma^{(3)}(\alpha_i)}{\Gamma(\alpha_i)} - 2\frac{\Gamma^{(2)}(\alpha_i)}{\Gamma(\alpha_i)} log\beta_i \\
&\quad + 2(\frac{\Gamma^{(1)}(\alpha_i)}{\Gamma(\alpha_i)})^2 log\beta_i - \frac{\Gamma^{(2)}(\alpha_i)}{\Gamma(\alpha_i)}\frac{\Gamma^{(1)}(\alpha_i)}{\Gamma(\alpha_i)} \\
Var[(log\lambda_{ij})^2] &= \frac{\Gamma^{(4)}(\alpha_i)}{\Gamma(\alpha_i)} - (\frac{\Gamma^{(2)}(\alpha_i)}{\Gamma(\alpha_i)})^2 - 4[\frac{\Gamma^{(3)}(\alpha_i)}{\Gamma(\alpha_i)} \\
&\quad - 4\frac{\Gamma^{(2)}(\alpha_i)}{\Gamma(\alpha_i)}\frac{\Gamma^{(1)}(\alpha_i)}{\Gamma(\alpha_i)}] log\beta_i \\
&\quad + 4[\frac{\Gamma^{(2)}(\alpha_i)}{\Gamma(\alpha_i)} - (\frac{\Gamma^{(1)}(\alpha_i)}{\Gamma(\alpha_i)})^2](log\beta_i)^2
\end{aligned}
$$

we can see the value of V_{ij} is determined by α_i , β_i and σ_i. Empirically we can assume α_i and β_i satisfying $tr([V'^{1/2}V_{ij}^{1/2}) < A$ for all i,j and some A. Then when I goes to infinity, denote $\lim \frac{\sum_i \sum_j V_{ij}}{IJ} = V$. by special case of Lindeberg-Feller central limit theorem stated in (Demidenko, 2013), $\sqrt{IJ}h(\theta) \xrightarrow{d} N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, V)$.

Proof of Lemma 4.6. From the assumptions we have $\alpha_i > 1$. From delta method:

$$
log(y_{ij}) \ - logN_{ij} \ = log(\lambda_{ij}) - \frac{1}{2\lambda_{ij}N_{ij}} + o(\frac{1}{\lambda_{ij}N_{ij}})
$$

$$
E\left(\frac{1}{y_{ij}+1}\right) = E\left[\frac{1}{\lambda_{ij}N_{ij}}(1 - e^{-N_{ij}\lambda_{ij}})\right] .
$$

From LLN,

$$\begin{aligned}
\Sigma_i\Sigma_j l_{ij} &\to \Sigma_i\Sigma_j E\left(log\frac{y_{ij}}{N_{ij}} + \frac{1}{2\,(y_{ij}+1)}\right) \\
&\approx \Sigma_i\Sigma_j[log(\lambda_{ij}) - \frac{e^{-N_{ij}\lambda_{ij}}}{2\lambda_{ij}N_{ij}}] \\
&\approx \Sigma_i J[\psi(\alpha_i) - log(\beta_i)] \\
\Sigma_i\Sigma_j\frac{(l_{ij}-\bar{l}_{i.})^2}{J-1} &\to \Sigma_i\frac{\Sigma_j var(l_{ij})}{J} \\
\Sigma_i\Sigma_j(\frac{1}{y_{ij}+1}) &\to \Sigma_i\Sigma_j E(\frac{1}{y_{ij}+1}) \to \Sigma_i\Sigma_j E(\frac{1}{y_{ij}+1}|\lambda_{ij}) \\
&\approx \Sigma_i\Sigma_j(\frac{1}{N_{ij}\lambda_{ij}})\,.
\end{aligned}$$

From delta method:

$$\begin{aligned}
&var(l_{ij}) \approx [E(\frac{1}{N_{ij}\lambda_{ij}}) + \psi_1(\alpha_i)] \\
&\Sigma_i\Sigma_j(l_{ij}-a)^2 \to \Sigma_i\Sigma_j E(l_{ij}-a)^2 = \Sigma_i\Sigma_j[E(l_{ij}-a)]^2 + \Sigma_i\Sigma_j var(l_{ij}) \\
&= \Sigma_i J[\psi(\alpha_i) - log(\beta_i) - a]^2 + \Sigma_i J\psi_1(\alpha_i) + \Sigma_i\Sigma_j\frac{1}{N_{ij}\lambda_{ij}} \\
&\Sigma_i\Sigma_j\frac{(x_{ij}-\bar{x}_{i.})^2}{J-1} \to \Sigma_i\frac{\psi_i(\alpha_i)}{b^2} + \Sigma_i\sigma_i^2\,.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\Sigma_i\Sigma_j(\frac{l_{ij}-a}{b})^2 + \Sigma_i\Sigma_j\frac{J(x_{ij}-\bar{x}_{i.})^2}{J-1} \\
&\approx \Sigma_i\Sigma_j\frac{[\psi(\alpha_i) - log(\beta_i) - a]^2 + \psi_1(\alpha_i)}{b^2} + \Sigma_i\Sigma_j\frac{\psi_1(\alpha_i) + \frac{1}{N_{ij}\lambda_{ij}}}{b^2} + \Sigma_i\Sigma_j\sigma_i^2 \\
&\to \Sigma_i\Sigma_j\frac{[\psi(\alpha_i) - log(\beta_i) - a]^2 + \psi_1(\alpha_i)}{b^2} + \Sigma_i\Sigma_j\frac{\psi_1(\alpha_i) + E[\frac{1}{N_{ij}\lambda_{ij}}]}{b^2} + \Sigma_i\Sigma_j\sigma_i^2 \\
&\approx \Sigma_i\Sigma_j\frac{[\psi(\alpha_i) - log(\beta_i) - a]^2 + \psi_1(\alpha_i)}{b^2} + \Sigma_i\Sigma_j\frac{J(l_{ij}-\bar{l}_{i.})^2}{b^2(J-1)} + \Sigma_i\Sigma_j\sigma_i^2\,.
\end{aligned}$$

That is

$$\Sigma_i\Sigma_j(\frac{l_{ij}-a}{b})^2+\Sigma_i\Sigma_j\frac{J(x_{ij}-\bar{x}_{i.})^2}{J-1}-\Sigma_i\Sigma_j\frac{J(l_{ij}-\bar{l}_{i.})^2}{b^2(J-1)}$$
$$\approx\Sigma_i\Sigma_j\frac{[\psi(\alpha_i)-log(\beta_i)-a]^2+\psi_1(\alpha_i)}{b^2}+\Sigma_i\Sigma_j\sigma_i^2\,.$$

Proof of Lemma 4.7. The first two approximation can be easily derived from Lemma 4.2. The third one follows

$$\begin{aligned}
\Sigma_i\bar{x}_{i.}l_i-I\bar{x}_{..}\bar{l}_{.}&\approx\ \Sigma_iE(\bar{x}_{i.}l_i)-IE(\bar{x}_{i.})E(l_i)\\
&\approx\Sigma_iE[E(\bar{x}_{i.}l_i|(\mu_i,log(\lambda_i))]-IE(\bar{x}_{i.})E(l_i)\\
&\approx\Sigma_iE[E[(\mu_i+\bar{\delta}_{i.})l_i|(\mu_i,log(\lambda_i))]]-IE(\bar{x}_{i.})E(l_i)\\
&\approx\Sigma_iE[E[\mu_il_i|(\mu_i,log(\lambda_i))]]-I\xi(a+b\xi)\\
&=\Sigma_iE[\mu_ilog(\lambda_i)]-I\xi(a+b\xi)\approx I\rho\sigma_x\sigma_y\,.
\end{aligned}$$

Proof of Lemma 4.8. The first two approximation can be easily derived from Lemma 4.6. The third one follows

$$\begin{aligned}
\Sigma_i\Sigma_jx_{ij}l_{ij}-IJ\bar{x}_{..}\bar{l}_{..}&\approx\ \Sigma_i\Sigma_jE(x_{ij}l_{ij})-IJ\bar{x}_{..}\bar{l}_{..}\\
&\approx\Sigma_i\Sigma_jE[E(x_{ij}l_{ij}|(\mu_{ij},log(\lambda_{ij}))]-IJ\xi(a+b\xi)\\
&\approx\Sigma_i\Sigma_jE[E[(\mu_{ij}+\delta_{ij})l_{ij}|(\mu_{ij},log(\lambda_{ij}))]]-IJ\xi(a+b\xi)\\
&\approx\Sigma_i\Sigma_jE[\mu_{ij}log(\lambda_{ij})]-I\xi(a+b\xi)\\
&\approx\Sigma_i\Sigma_jE[E[\mu_{ij}log(\lambda_{ij})|(\mu_i,log(\lambda_i))]]-I\xi(a+b\xi)\\
&\approx\Sigma_i\Sigma_jE[\mu_ilog(\lambda_i)]-IJ\xi(a+b\xi)\\
&=IJ\rho\sigma_x\sigma_y\,.
\end{aligned}$$

7.2 Bioinformatics background

Measurement of gene expression. The genetic information contained in genes influences on cells by the process of gene product synthesis named gene expression. Such product is often protein but sometimes is functional RNA. The essential step in the process is translation DNA to RNA molecules, therefore the technical approach to quantify gene expression is to measure the abundance of RNA molecules extracted from cell sample. Gene expression is the fundamental level of biological response to different phenotype

and treatment, hence detecting genes that are differentially expressed within certain experimental design based on measurements has been one primary topic in Bioinformatics researches.

Microarray. Microarray has become one of the major platform to measure genome-wide gene expressions. A Microarray is a glass plate where spots locate in an orderly fashion (Babu, 2004). Each spot contains a probe consisted of amplification of certain DNA molecules that bend to the RNA transcribed by a gene. To measure the gene expression in a sample, RNA molecules are extracted and convert to complementary DNAs (cDNA) with reverse transcription enzyme. Obtained cDNA molecules are dyed with red or blue and then hybridize with the probes in the glass plate. Genes with higher expression intensity tend to has more RNA transcripts in the sample, hence the spots of corresponding probes attract more dye and present brighter color. After this hybridization, the colored glass are scanned and the color intensity performs as a quantification of gene expression.

RNA-Seq. RNA-Seq is another appealing measurement platform quantifying thousands of genes' expression simultaneously based on sequencing technology. Similar to Microarray, RNA extracted from samples are covert to cDNA with reverse transcription. And long cDNA strand are shredded to millions of short fragments and one or both ends of the fragments are sequenced. With the sequence information, fragments' ends can be aligned to a reference genome and each aligned fragment end is called a aligned read. The counts of reads from gene regions quantifies corresponding gene expressions.

qRT-PCR. qRT-PCR is regarded as the most reliable platform to measure gene expression, the principle is to use color label molecules to monitor the polymerase chain reaction of cDNA obtained from reverse transcription of RNA. Despite the technical advantages of qRT-PCR, the throughput is low and measuring genes in a whole genome on qRT-PCR is not attractive in time and costs.

7.3 Supplementary methods

Mapping simulated gene to real gene. In simulating the RNA-Seq true intensity for genes: Given a vector of true intensities t generated from the

proposed Γ distribution, let $t' = log(t) + N(0, 0.04)$ and the rank R is set as the rank of t'. The a simulated gene with rank i in R are map to a real gene with rank i in the real data.

Exclude remaining zero counts. In random model, because each replicates need to be included individually in the estimator (22). Therefore it loses too much information to remove all genes with 0 count among replicates. Instead, when compute any summation, log(0) is replaced by the average of that summation excluding log(0). For example, $\Sigma_i \Sigma_j l_{ij}$ becomes $\Sigma_i \Sigma_j \frac{l_{ij}\phi(l_{ij}>-\infty)}{\Sigma_i \Sigma_j \phi(l_{ij}>-\infty)}$, where $\phi(\cdot)$ is the indicator function.

Balancing experimental design in Microbiome data. Due to quality control, two replicates from one subject at the same region are filtered out, which makes its sample mean on each region has a greater variance than other subject. To balance the design, additional replicates from other subjects are removed and each subjects has 8 replicates at the end.

Bibliography

Allen, M. J. and Yen, W. M. (2001). *Introduction to measurement theory*. Waveland Press.

Altman, D. G. and Bland, J. M. (1983). Measurement in medicine: the analysis of method comparison studies. *The statistician*, pages 307–317.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biol*, 11(10):R106.

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., and Robinson, M. D. (2013). Count-based differential expression analysis of rna sequencing data using r and bioconductor. *Nature protocols*, 8(9):1765–1786.

Anders, S., Pyl, P. T., and Huber, W. (2014). Htseq–a python framework to work with high-throughput sequencing data. *bioRxiv*.

Atchadé, Y. F. (2011). A computational framework for empirical bayes inference. *Statistics and computing*, 21(4):463–473.

Babu, M. M. (2004). Introduction to microarray data analysis. *Computational Genomics: Theory and Application*, pages 225–249.

Barnett, V. (1970). Fitting straight lines-the linear functional relationship with replicated observations. *Applied Statistics*, pages 135–144.

Beane, J., Vick, J., Schembri, F., Anderlind, C., Gower, A., Campbell, J., Luo, L., Zhang, X. H., Xiao, J., Alekseyev, Y. O., et al. (2011). Characterizing the impact of smoking and lung cancer on the airway transcriptome using rna-seq. *Cancer prevention research*, 4(6):803–817.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

Bollen, K. A. (1998). *Structural equation models*. Wiley Online Library.

Brace, R. A. (1977). Fitting straight lines to experimental data. *Am. J. Physiol*, 233(3):R94–R99.

Carroll, R. J. (1998). Measurement error in epidemiologic studies. *Encyclopedia of biostatistics*.

Carroll, R. J., Spiegelman, C. H., Lan, K. G., Bailey, K. T., and Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika*, 71(1):19–25.

Castilho, M. V. (2004). A comparison of statistical techniques for detecting analytical bias in geoanalysis. *Geostandards and Geoanalytical research*, 28(2):277–290.

Chakravorty, S., Helb, D., Burday, M., Connell, N., and Alland, D. (2007). A detailed analysis of 16s ribosomal rna gene segments for the diagnosis of pathogenic bacteria. *Journal of microbiological methods*, 69(2):330–339.

Chan, L. K. and Mak, T. K. (1979). Maximum likelihood estimation of a linear structural relationship with replication. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 263–268.

Cheetham, S., Tang, M., Mesak, F., Kennecke, H., Owen, D., and Tai, I. (2008). Sparc promoter hypermethylation in colorectal cancers can be reversed by 5-aza-2 deoxycytidine to increase sparc expression and improve therapy response. *British journal of cancer*, 98(11):1810–1819.

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., and Liu, C. (2011). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one*, 6(2):e17238.

Chen, X., Hong, H., and Nekipelov, D. (2007). Measurement error models. *Prepared for the Journal of Economic Literature. www. stanford. edu/˜ doubleh/eco273B/surveyjan27chenhandenis-07. pdf*.

Cornbleet, P. J. and Gochman, N. (1979). Incorrect least-squares regression coefficients in method-comparison analysis. *Clinical chemistry*, 25(3):432–438.

Cox, D. (1993). Unbiased estimating equations derived from statistics that are functions of a parameter. *Biometrika*, 80(4):905–909.

Cragg, J. G. (1997). Using higher moments to estimate the simple errors-in-variables model. *Rand Journal of Economics*, pages S71–S91.

Demidenko, E. (2013). *Mixed Models: Theory and Applications with R*. John Wiley & Sons.

Durbin, B. P., Hardin, J. S., Hawkins, D. M., and Rocke, D. M. (2002). A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18(suppl 1):S105–S110.

Flicek, P., Aken, B. L., Beal, K., Ballester, B., Cccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T. A., Dyer, S. C., Eyre, T., and et al. (2008). Ensembl 2008. *Nucleic Acids Research*, 36(Database-Issue):707–714.

Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., Menzel, C., Chen, W., Li, Y., Zeng, R., et al. (2009). Estimating accuracy of rna-seq and microarrays with proteomics. *BMC genomics*, 10(1):161.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affyanalysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315.

Gillard, J. (2006). An historical overview of linear regression with errors in both variables. *Cardiff University School of Mathematics Technical Report*.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Hardcastle, T. J. and Kelly, K. A. (2010). bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, 11(1):422.

Hox, J. and Bechger, T. (1998). An introduction to structural equation modelling. *Family Science Review*, 11(354-373).

Huang, E., Zhu, W., Dhundale, A., Bahou, W., and Gnatenko, D. (2013). Platelet genetic biomarker quantification: comparison of fluorescent microspheres. *Thromb Haemost*, 109(2):337–46.

Huber, W., Von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and

to the quantification of differential expression. *Bioinformatics*, 18(suppl 1):S96–S104.

Hyman, E., Kauraniemi, P., Hautaniemi, S., Wolf, M., Mousses, S., Rozenblum, E., Ringnér, M., Sauter, G., Monni, O., Elkahloun, A., et al. (2002). Impact of dna amplification on gene expression patterns in breast cancer. *Cancer research*, 62(21):6240–6245.

Kendall, M. G. et al. (1946). The advanced theory of statistics. *The advanced theory of statistics.*, (2nd Ed).

Kukush, A., Schneeweis, H., and Wolf, R. (2004). Three estimators for the poisson regression model with measurement errors. *Statistical Papers*, 45(3):351–368.

Kvam, V. M., Liu, P., and Si, Y. (2012). A comparison of statistical methods for detecting differentially expressed genes from rna-seq data. *American journal of botany*, 99(2):248–256.

Li, J. and Tibshirani, R. (2013). Finding consistent patterns: A nonparametric approach for identifying differential expression in rna-seq data. *Statistical methods in medical research*, 22(5):519–536.

Liang, H. (2000). *Errors-in-Variables Models*. Springer.

Lindley, D. V. (1947). Regression lines and the linear functional relationship. *Supplement to the Journal of the Royal Statistical Society*, pages 218–244.

Lindsay, B. (1982). Conditional score functions: some optimality results. *Biometrika*, 69(3):503–512.

Linnet, K. (1993). Evaluation of regression procedures for methods comparison studies. *CLINICAL CHEMISTRY-WASHINGTON-*, 39:424–424.

Ludbrook, J. (1997). Special article comparing methods of measurement. *Clinical and Experimental Pharmacology and Physiology*, 24(2):193–203.

Ludbrook, J. (2010). Linear regression analysis for comparing two measurers or methods of measurement: but which regression? *Clinical and Experimental Pharmacology and Physiology*, 37(7):692–699.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517.

Martin Bland, J. and Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476):307–310.

McCullagh, P. and Nelder, J. A. (1989). Generalized linear models.

McDonald, J. H. (2009). *Handbook of biological statistics*, volume 2. Sparky House Publishing Baltimore, MD.

Mooney, M., Bond, J., Monks, N., Eugster, E., Cherba, D., Berlinski, P., Kamerling, S., Marotti, K., Simpson, H., Rusk, T., et al. (2013). Comparative rna-seq and microarray analysis of gene expression changes in b-cell lymphomas of canis familiaris. *PloS one*, 8(4):e61088.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628.

Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika*, 77(1):127–137.

Osborne, C. (1991). Statistical calibration: a review. *International Statistical Review/Revue Internationale de Statistique*, pages 309–336.

Passing, H. and Bablok, W. (1983). A new biometrical procedure for testing the equality of measurements from two different analytical methods. application of linear regression procedures for method comparison studies in clinical chemistry, part i. *Clinical Chemistry and Laboratory Medicine*, 21(11):709–720.

Rau, A., Gallopin, M., Celeux, G., and Jaffrézic, F. (2013). Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics*, 29(17):2146–2152.

Ren, S., Peng, Z., Mao, J.-H., Yu, Y., Yin, C., Gao, X., Cui, Z., Zhang, J., Yi, K., Xu, W., et al. (2012). Rna-seq analysis of prostate cancer in the chinese

population identifies recurrent gene fusions, cancer-associated long noncoding rnas and aberrant alternative splicings. *Cell research*, 22(5):806–821.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Rocke, D. M. and Durbin, B. (2001). A model for measurement error for gene expression arrays. *Journal of computational biology*, 8(6):557–569.

Salzman, J., Jiang, H., and Wong, W. H. (2011). Statistical modeling of rna-seq data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 26(1).

Scott, E. L. (1950). Note on consistent estimates of the linear structural relation between two variables. *The Annals of Mathematical Statistics*, pages 284–288.

Shwetha, S., Gouthamchandra, K., Chandra, M., Ravishankar, B., Khaja, M., and Das, S. (2013). Circulating mirna profile in hcv infected serum: novel insight into pathogenesis. *Scientific reports*, 3.

Sinicropi, D., Qu, K., Collin, F., Crager, M., Liu, M.-L., Pelham, R. J., Pho, M., Dei Rossi, A., Jeong, J., Scott, A., et al. (2012). Whole transcriptome rna-seq analysis of breast cancer recurrence risk using formalin-fixed paraffin-embedded tumor tissue. *PloS one*, 7(7):e40092.

Smyth, G. K. et al. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3(1):3.

Ståhlberg, A., Kubista, M., and Pfaffl, M. (2004). Comparison of reverse transcriptases in gene expression analysis. *Clinical chemistry*, 50(9):1678–1680.

Stefanski, L. (1988). Measurement errors in generalized linear model explanatory variables. In *3rd International Workshop on Statistical Modelling, Vienna*, pages 1–40.

Sun, Z., Asmann, Y. W., Nair, A., Zhang, Y., Wang, L., Kalari, K. R., Bhagwate, A. V., Baker, T. R., Carr, J. M., Kocher, J.-P. A., et al. (2013).

Impact of library preparation on downstream analysis and interpretation of rna-seq data: comparison between illumina polya and nugen ovation protocol. *PloS one*, 8(8):e71745.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3):562–578.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, 11(3):284–300.

Wong, M. (1989). Likelihood estimation of a simple linear regression model when both variables have error. *Biometrika*, 76(1):141–148.

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

Xu, P., Liu, J., Zeng, W., and Shen, Y. (2014). Effects of errors-in-variables on weighted least squares estimation. *Journal of Geodesy*, pages 1–12.

Xu, X., Zhang, Y., Williams, J., Antoniou, E., McCombie, W. R., Wu, S., Zhu, W., Davidson, N. O., Denoya, P., and 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.

Zamar, R. H. (1989). Robust estimation in the errors-in-variables model. *Biometrika*, 76(1):149–160.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. *PloS one*, 9(1):e78644.